Genome-Scale Quantitative Biology of *Arabidopsis thaliana*

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Marc William Schmid

von

Muttenz BL und Schaffhausen SH

Promotionskomitee

Prof. Dr. Ueli Grossniklaus (Vorsitz und Leitung der Dissertation)

Dr. Jörg Becker

Prof. Dr. Mark Robinson

Zürich, 2015

Abstract

The formation of gametes is a fundamental process for sexual reproduction. In contrast to animals, where haploid gametes are formed directly through meiosis, plants first form haploid spores. These spores typically give rise to multicellular organisms (gametophytes), which mature through regular mitotic divisions and finally harbour the gametes. In flowering plants, such as *Arabidopsis thaliana*, the gametophytes are reduced to only a few cells: the three-cellular pollen grain (male) and a seven-celled embryo sac (female). The seven cells of the embryo sac represent four very distinct cell types, the two female gametes, the egg and the central cell, the synergids involved in pollen tube reception, and the antipodals. The identity of these cell types is specified just after cellularization of a single cell containing eight nuclei (syncytium). Oriented mitosis, nuclear migration, and the position of the nuclei prior to cellularization were shown to be crucial for cell-fate aquisition. Little is known about the molecular mechanisms underlying these patterning processes. We hypothesized that specific subcellular localization of mRNA may be involved. Thus, we isolated the two opposing poles of the syncytium using laser-assisted microdissection (LAM) and compared their mRNA content by RNA sequencing (RNA-Seq). We found that a substantial number of transcripts (615) are preferentially localized at only one pole of the syncytium. Some of them seem to act as cell-fate determinants, whereas others may control the localization of the proteins they encode for. The novel insights from this study can aid the in-depth investigation of genes and regulatory mechanisms involved in gametogenesis.

The isolation and profiling of the two cell halves and the subsequent data analysis required establishing the combination of LAM and RNA-Seq, and the development of software tools for data analysis. During the course of my PhD studies, I also contributed with programming, data analysis, and interpretation to various collaborative projects addressing many aspects of plant biology from evolution of sex-specific genes to nuclear architecture. Consequently, the work presented in this thesis comprises several research topics, experimental procedures, and data analysis approaches. The unifying theme of this thesis is therefore the attempt to understand various complex aspects of plant biology with the help of quantitative large-scale data.

Zusammenfassung

Die Bildung von Gameten ist ein zentraler Prozess der sexuellen Fortpflanzen. Im Unterschied zu Tieren, welche die haploiden Gameten direkt durch die Meiose bilden, werden bei Pflanzen zuerst haploide Sporen gebildet. Diese Sporen entwickeln sich dann durch reguläre, mitotische Zellteilungen zu multizellulären Organismen (Gametophyten), welche die Gameten enthalten. Die Gametophyten der Blütenpflanzen, wie zum Beispiel diejenigen von *Arabidopsis thaliana*, sind stark reduziert und bestehen aus nur wenigen Zellen. Die männlichen und weiblichen Gametophyten bestehen jeweils aus nur drei und sieben Zellen. Die sieben Zellen des weiblichen Gametophyten repräsentieren jedoch vier sehr unterschiedliche Zelltypen: Die zwei Gameten (Eizelle und Zentralzelle), die Synergiden, welche die Pollenschläuche anlocken und empfangen, und die Antipoden. Die Differenzierung dieser vier Zelltypen erfolgt vermutlich während oder kurz nach der Zellularisierung einer einzigen Zelle mit mehreren Zellkernen (Syncytium). Dabei spielt die Position der Zellkerne eine wichtige Rolle. Diese wird durch orientierte Kernteilungen und zielgerichtete Translokation der Kerne etabliert. Bislang ist weitgehend unbekannt, wie diese Prozesse reguliert werden. Wir vermuteten, dass eine ungleichmässige Verteilung von Transkripten (mRNA) innerhalb des Syncytiums wichtig sein könnte. Um unsere Hypothese zu testen, isolierten wir die zwei gegenüberliegenden Zellhälften des Syncytiums mithilfe von lasergestützter Mikrodissektion und verglichen deren mRNA Gehalt mittels RNA-Seq. Dabei identifizierten wir 615 Gene, deren Transkripte vorwiegend in einer der beiden Zellhälften angereichert waren. Einige davon agieren vermutlich als Determinanten für die Spezifizierung eines bestimmten Zelltypen. Andere wiederum sind vermutlich spezifisch lokalisiert um die Lokalisation des von ihnen kodierten Proteins zu kontrollieren. Die Resultate der Studie können eine wichtige Basis zur weiteren Charakterisierung von Genen und regulatorischen Mechanismen, welche in der Gametogenese involviert sind, bilden.

Um die Zellhälften zu isolieren, die Transkripte zu sequenzieren und die Daten zu analysieren, mussten wir zuerst die Methode etablieren und geeignete Programme für die Datenanalyse entwickeln. Während meines Doktorats habe ich darüber hinaus auch entscheidend zur Datenanalyse und Dateninterpretation in mehreren kollaborativen Projekten beigetragen. Die Fragestellungen dieser Studien reichen von der Evolution geschlechtsspezifischer Gene zur Architektur des Genoms innerhalb des Zellkerns. Die hier präsentiere Arbeit umfasst deshalb ein breites Spektrum an Themen, experimentellen Methoden und analytischen Vorgehensweisen. Das umfassende Thema dieser Dissertation ist demnach der Versuch, komplexe biologische Vorgänge mithilfe von grossen, quantitativen Datensätzen zu verstehen.

Acknowledgements

I would like to thank:

Ueli Grossniklaus for being a friend, his support, and the freedom I had during my PhD thesis.

Anja Schmidt for all the collaborations, the scientific as well as non-scientific discussions, and her comments on some of the manuscripts.

Stefan Grob for the interesting discussions on chromosomal architecture and the great time we had at and after work.

Anja Herrmann and Daniela Guthörl for their experimental work, which gave me the time to finalize Rcount and HiCdat.

Christian Heichinger for fruitful discussions and entertaining coffee breaks.

Hannes Vogler and Afif Hedhly for sharing a countless number of beers.

The members of the lab for creating a friendly atmosphere with lots of fun and parties.

Most importantly – my family: Diana, Alissa, Hanni, Bernhard, Barbara, Rahel, Roland, and Peter.

Contents

1 Prologue

All multicellular organisms develop from one initial cell. In humans and most other animals, this initial cell is the zygote, which is formed through fertilization of the egg cell. Plants and fungi follow the same principle, but in addition to the organism developing from the zygote, they go through a second multicellular generation in their life cycle. In the life cycle of plants, the two multicellular organisms are the sporophyte and the gametophyte. The mature sporophyte produces spores through reductive cell division (meiosis). These spores give rise to the gametophytes, which generate the gametes through regular cell division (mitosis). Fusion of a male and female gamete results in the formation of a zygote. The zygote finally develops into a new sporophyte, thereby closing the life cycle. Variations in the embodiment of this basic life cycle over the course of evolution contribute to the diversity of species observed today. Whereas the gametophytes represent the dominant life form of bryophytes, they are reduced to only few cells completely dependent on the sporophyte in spermatophytes. Even though small, the gametophytes of flowering plants are of outstanding importance for mankind. For example, in-depth understanding of the processes underlying female gametogenesis in flowering plants bears great potential for further crop improvements and simplification of breeding efforts (chapter 2).

The aim of my PhD thesis was to test, whether transcripts exhibit polarized localization during the syncytial stages of female gametophyte development in *Arabidopsis thaliana*, and if, which role this process may play (chapter 3). To facilitate the isolation and the profiling of the cell halves, I first established the combination of laser-assisted microdissection (LAM) followed by high-throughput RNA sequencing (RNA-Seq). Preliminary data were generated during my Master thesis, where I performed a pilot RNA-seq experiment on the central cell and the two cell halves. However, in-depth analysis of these data at the beginning of my PhD studies revealed a lack of suitable software for RNA-Seq data pre-processing (chapter 4). Commonly used count algorithms did not address the problem of reads aligning with multiple locations in the genome (multireads) or reads aligning with positions where two or more genes overlap (ambiguous reads). I therefore developed Rcount, a tool for RNA-Seq data pre-processing specifically addressing both problems (chapter 5). During the course of my PhD studies, I participated in various collaborative projects where I contributed with programming, data analysis and interpretation, and statistical consulting (chapter 8). The two most significant contributions were in the projects studying the nuclear architecture of *Arabidopsis thaliana* from Stefan Grob (section 8.1 and 8.2), where my work on data analysis and software development (e.g., chapter 6) was decisive for our rapid success. Another substantial contribution was the development of a novel statistical analysis method for whole genome bisulfite sequencing data by using multifactorial linear models (Christian Heichinger, unpublished).

All chapters of this thesis are written as independent manuscripts. Published manuscripts are included as layouted by the publisher. Supplemental material can either be found online (published articles) or at the end of the manuscript (unpublished manuscripts, except large zip-archives and txt-tables provided electronically). Rcount and HiCdat are freely available on github.com/MWSchmid. The contribution of other authors to my work is acknowledged on the title pages preceding the manuscripts. Likewise, my contributions to the manuscripts of other main authors are given. Chapters 2 to 6 comprise five manuscripts in which I was the project leader and main contributor to the text, experiments, data analysis, and software design and development. Chapter 8 contains published research articles and reviews, to which I contributed. Unpublished work to which I contributed is not included in this thesis.

Results are discussed in detail at the end of each manuscript. In addition, chapter 2 discusses the future directions and perspectives of the microgenomics (i.e., cell type-specific systems biology) approach to plant developmental biology, specifically the research on female gametogenesis. Given that this discussion is very broad and not directly applicable to the main topics covered in this thesis, some immediate future perspectives are presented in the epilogue (chapter 7).

2 The female gametophyte: An emerging model for cell type-specific systems biology in plant development

The following manuscript is a review intended to contribute to the research topic "Plant Single Cell Type Systems Biology" in Frontiers (frontiersin.org). I wrote the abstract and the sections 1, 3, and 4. Anja Schmidt contributed with section 2 ("methods for aquisition of large-scale quantitative data of specific cell types" until "systems biology approach toward plant development"). All sections were critically read and modified by Anja Schmidt and myself. We both designed the figures and the tables. I generated all figures and Anja Schmidt made the tables.

The female gametophyte: An emerging model for cell type-specific systems biology in plant development

Marc W. Schmid[1,†,*], Anja Schmidt[1,†,*], Ueli Grossniklaus[1]
1 Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zürich, Switzerland
† These authors contributed equally to this work.
*** E-mail: marcschmid@gmx.ch, aschmidt@botinst.uzh.ch**

Abstract

Systems Biology, a holistic approach describing a system emerging from the interactions of its molecular components, critically depends on accurate qualitative determination and quantitative measurements of these components. Development and improvement of large-scale profiling methods ("omics") now facilitate comprehensive measurements of many relevant molecules. For multicellular organisms like plants, the complexity of the system is augmented by the presence of specialized cell types and organs, and a complex interplay within and between them. Cell type-specific analyses are therefore crucial for the understanding of developmental processes and environmental responses. This review first gives an overview of current methods used for large-scale profiling of specific cell types exemplified by recent advances in plant biology. The focus then lies on suitable model systems to study plant development and cell type specification. We thereby introduce the female gametophyte of flowering plants as an ideal model to study fundamental developmental processes. Moreover, the female reproductive lineage is of importance for the emergence of evolutionary novelties such as an unequal parental contribution to the tissue nurturing the embryo or the clonal production of seeds by asexual reproduction (apomixis). Understanding these processes is not only interesting from a developmental or evolutionary perspective, but bears great potential for further crop improvement and simplification of breeding efforts. We finally highlight novel methods, which are already available or which will likely soon facilitate large-scale profiling of the specific cell types of the female gametophytes in both, model as well as non-model species. We conclude that it may take only few years from now until an evolutionary systems biology approach toward female gametogenesis may decipher some of its most interesting and economically most valuable processes.

Systems biology: an integrated approach to model biological processes with large-scale quantitative data

Since the foundation of the Institute for Systems Biology in the year 2000 and the formal introduction of systems biology at the beginning of the 21st century [1,2], systems biology has been a steadily growing field of research. As an integrative approach, it is markedly different from the reductionistic approach generally used in molecular biology and genetics. Powered by the central dogma of biology, where a gene is transcribed to mRNA, which is then translated into a protein, molecular biology and genetics have successfully identified genes, their functions, and processes they are involved in. However, the implicit link of a gene to a certain function or a phenotype is a strong oversimplification of the underlying process. It thus frequently misses important interactions with other cellular or environmental factors (e.g., responses to environmental conditions like the temperature dependent phenotype of a mutant). In contrast, systems biology may be described as an attempt to quantitatively describe and understand the global behavior of a biological entity emerging from the interactions between its molecular components. Such comprehensive understanding would allow for prediction and modelling of the biological entity, its precise control and ultimately targeted manipulation of complex biological systems (reviewed in [2–7]).

Systems biology comprises and integrates experimental studies and large-scale data sets derived from high-throughput technologies ("omics"), such as transcriptomics (RNA profiling), proteomics (analysis of proteins), and metabolomics (profiling of metabolites). However, also epigenetic regulatory processes based on the modification of chromatin structure or DNA methylation (epigenomics), the translation of mRNAs to proteins (translatomics), complex formation of proteins with proteins or nucleic acids (interactomics), the investigation of protein modifications, e.g. phosphorylation important for regulation of their activity, and the transport of ions or metabolites (fluxomics) need to be taken into account to achieve a full picture of the dynamic processes of a cell or organism (reviewed by [8]). One of the most crucial aspects for systems biology is the comprehensiveness of the "omics" data [2]. For a given method this includes the number of items that can be measured at once (e.g., transcripts with transcriptomics). For the entire system, it is then important whether the relevant items (e.g., enzymes and metabolites) or processes (e.g., posttranslational modifications) can be accurately measured with a combination of certain methods. An additional level of complexity may be imposed by the requirement of a high spatial and/or temporal resolution. For a single, isolated cell, this can refer to specific organelles, subcellular compartments, certain domains of the plasma membrane, and the stage of the cell-cycle. For an unicellular organism like yeast, this may be augmented by studying the cell-to-cell variability within the population [9]. In multicellular organisms, each cell (type) has a specific function and position within

an organ. Its role and differentiation status may be influenced by local signals as well as systemic signals originating from other organs (e.g., hormones). In addition, the temporal coordinate expands to developmental stages of the organs or the life span of the organism.

Consequently, a complete understanding at the system level requires highly resolved quantitative spatio-temporal data on the individual components and their interactions, and the integration of the data into models. On one hand, integration of these data with computational methods can aid to characterize previously unknown components (e.g., genes) of a system, as exemplified for yeast [10]. Alternatively, the data may be used in a mathematical model describing the system and allowing for prediction of system behavior and hypothesis generation [11]. Finally, the integration of omics data, the formulation of mathematical models, the generation of hypotheses, and the experiments are interlinked and benefit from each other. A possible extension of systems biology is the use of interspecies comparisons to, for example, elucidate the extent to which genotypic variation translates into phenotypic differences [12]. Even broader, evolutionary systems biology may be recognized as an approach to describe and understand how biological systems are shaped by evolution and are steering it at the same time (reviewed in [13]).

Prior to the understanding of a complex organism composed of many different cell and tissue types, investigations of distinct cell types can lead to an understanding of basic processes governing cellular specification, identity and metabolism. To date, yeast (*S. cerevisiae*) is a widely used model system appreciated to be the currently best understood cell [14]. While evolutionary only distantly related, pathways in yeast have shown to have considerable similarities to the ones in plants, animals and humans [1]. In addition, yeast serves in production of food and pharmaceuticals. Due to its simplicity and its importance for biotechnology and biomedical research, yeast has shaped modern molecular biology to a great extent. Indeed, it has been a pioneering organism in systems biology (reviewed in [14–16]), starting from gene expression and regulatory networks discovered during early transcriptome studies and their integration with other genome wide data, over genetic interaction networks obtained by crossing thousands of mutant strains [17] and modeling of gene expression as quantitative trait (eQTL, [18]), to genome-wide metabolic models. However, given the unicellularity of yeast, it can hardly serve as a developmental model for complex multicellular animals and even less for plants. In plants, systems biology is less advanced due to several reasons, including the higher complexity of most plant genomes, gene families, and the multitude of primary and secondary metabolites, as well as the lack of suitable *in vitro* systems or cell lines for most plant tissue types or organs. Most efforts in plant research thus require *in vivo* experiments making the procedures generally more difficult and less suitable to high-throughput approaches. As a consequence, data generation can sometimes still be a severely limiting factor for plant sciences. On the other hand, the results are therefore of high relevance for the process under investigation.

Apart from this, substantial progress in the analysis of specific cell types in plants has been made over the last decade. Facilitated by advances in high-throughput profiling technologies and methods for the isolation of specific cell types, recent studied focussed on the analysis of specific cell types or even single cells (Figure 1). To investigate cell type-specific processes in higher plants, root hairs and trichomes have been used as models, both for their physiological importance and their decent accessibility at the epidermal surfaces (for details see below; [19–31]). In addition, starting from only few examples at the beginning of the 21st century [32], cell type-specific transcriptional profiling has become a robust and frequently used method. In the model plant *Arabidopsis thaliana*, novel insights into plant development and cellular response to environmental stimuli were for example gained through studies on individual cell types of the root, root hairs, trichomes, and guard cells, and by transcriptional profiling during male and female gametogenesis (reviewed in [33–35]). These examples clearly point out the importance of cell type-specific investigations for a detailed understanding of the differentiation processes and environmental responses of distinct cell types. However, depending on the cell type under investigation, the currently available methods for cell isolation may still be challenging, time-consuming, and limited to a subset of "omics" approaches (e.g., laser-assisted microdissection on rare cell types, [35]). While studies focusing on specific cell types, which can be isolated in quantities high enough for the full set of "omics" approaches, could serve as initial models for cell type-specific systems biology in plants [23], the ultimate goal must be that the full set of methods can be applied to any cell type of interest.

Methods for aquisition of large-scale quantitative data of specific cell types

Large-scale profiling of distinct cell types critically depends on the possibility to isolate these cells with sufficient purity and quantity as well as the sensitivity and accuracy of the profiling methods. Despite the rapid improvements of existing and development of novel tools for systems biology, the demand for fast and easily applicable methodologies for cell type-specific analyses are not yet satisfied. Further challenges are associated with the requirement for normalization and integration of different data types, and the increasing demand for platforms allowing storage and sharing of the rapidly growing amount of large-scale datasets (reviewed by [5,6,8,36]). In brief, three steps are generally of great importance for cell type-specific systems biology: (i) Isolation and purification of the specific cell type, (ii) profiling of a certain molecular compound, and (iii) data analysis, integration, storage and sharing. In the following sections, we present current methods to aquire large-scale quantitative data required for systems biology. The focus lies on methods allowing for genome-wide cell type-specific analyses and on representative examples. For a discussion on the computational challenges in systems biology, the reader is refered to several recent reviews [3–6, 36–40].

Methods for cell type-specific isolation

A restricted number of cell types in plants is exposed at the tissue surfaces and can be collected by abrasion or mechanical detachment. Depending on the species, relatively simple mechanical isolation procedures for trichomes and root hairs enabled a large spectrum of methods. Mechanical isolation of trichomes allowed for transcriptomics in various species as well as metabolomics (see [22] for an integrated database) and proteomics [25,27]. Another example for an exposed cell type are the root hairs, for which the relatively simple isolation procedures facilitated transcriptomics [24], proteomics [20, 26], and metabolomics [21]. Certain other cell types can further be isolated by tissue disruption followed by centrifugation based methods, or through mechanical preparation and manual microdissection. Examples include specific cell types from the male or female reproductive lineages, plant mesophyll cells and guard cells (reviewed by [28, 34, 35]). Proteomic profiling has, for example, been performed from *Brassica napus* guard cells and mesophyll cells that could be purified as protoplasts [41]. However, for most cell types, these methods are not applicable. Several methods for the isolation of specific cell types embedded in differentiated tissues have been established. Fluorescent activated cell sorting (FACS) can be used to sort fluorescent cells based on their light scattering characteristics and on fluorescence (reviewed by [42]). This method already allowed high resolution transcriptional profiling for different cell types of the *Arabidopsis thaliana* root, more recently proteomics [43], and metabolite mapping of the root cell and tissue types ([44]; reviewed by [45,46]). Similarly, fluorescent sorting of nuclei (FANS) has been established and, for example, used to isolate endosperm nuclei for profiling of RNA activity or epigenetic modifications [47,48]. Despite the great potential of FACS/FANS for plant cell type-specific systems biology, both approaches have certain limitations: They can only be applied if transgenic lines carrying cell type-specific fluorescent markers can be established, and they are thus not suitable for non-model species. In addition, depending on the tissue type, longer enzymatic incubations are required to digest the cell walls and to release the protoplasted cells prior to sorting [49]. Consequently, changes in, for example, the transcriptome or metabolome cannot fully be excluded. Alternatively, the INTACT method (isolation of nuclei in specific cell types) allows the isolation of nuclei expressing a biotinylated nuclear envelope protein by affinity purification with streptavidin-coated beads [50]. This method is suitable to study epigenetic modifications (DNA methylation of histone modifications) and to profile the RNA within the nucleus. To study actively translated mRNAs bound to ribosomes (translatome), small epitope tags can be fused to a ribosomal protein to allow immunopurification of the ribosomes containing the mRNAs with a method named TRAP (reviewed in [51]). Alternatively RNAs binding to RNA binding proteins involved in the formation of ribonucleoprotein (RNP) complexes can be profiled by immunoprecipitation of an epitope tagged protein (RIP; [51]). It has to be noted that the analyses of transcriptome and translatome abundance will not give the same results as not all mRNAs present in a cell are actively translated at a time. In

this respect profiling of mRNAs bound to ribosomes gives complementary results to transcriptome profiling as the readouts are closer to the synthesis of proteins ([51]). Similar to FACS and FANS, also INTACT, TRAP and cell type-specific RIP require the use of transgenic lines and pre-existing knowledge about cell type-specific promoters or markers.

An alternative method not requiring any molecular knowledge is laser-assisted microdissection (LAM). Plant tissues are thereby typically fixed and embedded in paraffin wax (reviewed in [34, 35]) or special resins [52, 53]. Thin sections of the tissues (typically between $6 - 10\,\mu$m) are subsequently mounted on metal framed plastic slides and used to isolate the cell types of interest after resolving the wax or resin and drying the tissues on the slides [53, 54]. The main constraints of the method is that harvesting sufficient material for downstream "omics" methods can be very time consuming. Furthermore, the suitability for single cell isolation depends on the optical resolution of the sectioned tissues and the visibility of the cell type of interest in addition to the physical properties of the laser beam in the instrument used [34]. Thus the time required for collecting enough material for one sample is largely dependent on the cell type of interest. So far, the applications of LAM for cell type-specific "omics" have been restricted to transcriptional profiling, e.g. to study cell type-specification in the female reproductive lineage in *Arabidopsis thaliana*, *Boechera gunnisoniana*, and *Hieracium praealtum* [53,55–58]. However, other applications, such as genome wide profiling of DNA methylation, are likely feasible (see below).

Methods for data acquisition

Transcriptomics

Transcriptome profiling encompasses the identification and quantification of all expressed RNA transcripts at a time (mRNA, tRNA, microRNA). However, due to the frequent use of oligo-dT priming during cDNA synthesis or transcriptome microarrays (i.e., covering only coding regions of the genome), many studies are restricted to mRNAs or a subset of mRNAs. Several types of microarrays were produced and extensively used for the analyses of gene expression in different plant species, including the model plant *Arabidopsis thaliana* and different important crop species like maize, rice, and barley (reviewed in [8]). The Affymetrix ATH1 GeneCHIP (`www.affymetrix.com`), the most popular microarray for *Arabidopsis thaliana*, has for example been used to profile a large variety of different tissue types (e.g., [59]), specific cell types of the root isolated through FACS [60], and specific cell types of the male and female reproductive lineages (reviewed in [34]). In addition to well established tools for data analysis, the wealth of publicly available datasets generated on the same platform makes commonly used microarrays a very valuable tool for systems biology [6].

Apart from microarrays, several platforms for Next Generation Sequencing (NGS) have been developed over the last years and are now routinely used for transcriptional profiling (RNA-Seq; see [61] for a review on NGS platforms). RNA-Seq has several advantages as compared to the use of microarrays, including a higher dynamical range, higher sensitivity, and the whole genome coverage allowing for identification of previously unknown transcripts and splice variants (reviewed in [34]). A major advantage is the applicability to non-model species, either through *de novo* assembly of the short reads into transcripts or by the use of a reference transcriptome either produced separately or taken from a public database (e.g., the ongoing effort to sequence 1'000 plant transcriptomes, `www.onekp.com`). Examples for such an approach are the central cells of *Arabidopsis thaliana* and cells of the female reproductive lineage in *Hieracium praealtum*, and *Boechera gunnisoniana* [53, 57, 58]. Several tools for RNA-Seq data analysis are available (see [8] for a selection of software tools, and [62] for a count tool addressing the problem of reads aligning at multiple locations in the genome, or reads aligning at positions where two or more genes overlap; Rcount). Current challenges are the increasing demand for standardized annotations of datasets and the development of computational methods allowing the integration of data from different studies using different methods and platforms. In perspective, the integration of data from different species will be of great value for plant systems biology allowing to gain insights into conserved common regulatory mechanisms, environmental adaptations and evolutionary changes.

Proteomics

Aside the analysis of gene expression (transcriptomics) and actively translated mRNAs (translatomics), the investigation of proteins (proteomics) and protein modifications (e.g., phosphoproteomics) add additional levels of complexity. From a systems biology perspective the aim would be the combination of cell type-specific proteomics with transcriptomics and metabolomics to elucidate and model regulatory networks (reviewed in [28]). In the beginning of proteomics, 2D gel electrophoresis was frequently used for separation of the proteins in a sample and to identify spots representing proteins differentially represented in two samples (reviewed by [63]). However, the protein or protein mixture in one spot could only be identified by excising the spot and analysis with mass spectrometry (MS). To date, proteomics is largely dependent on the use of mass spectrometry methods (MS). Typically, proteins are first digested with trypsin and subsequently either analyzed directly by MS or first separated by chromatography before MS. MS methods have been greatly improved with the development of soft ionization methods like electrospray ionization (ESI) in solution (typically aqueous or organic solvents) or matrix assisted laser desorption ionization (MALDI, [63, 64]). By both methods, intact, gas phase ions are generated that are introduced in mass analyzers and sorted depending on their mass to charge ratio, e.g. using time-of-flight (TOF, [64]; for a recent summary of mass analyzers see [65]; for a description of Orbitrap mass analysers see [66]). However, detection

based on peptide mass to charge ratios is largely qualitative and can only be used for quantification in two or more samples acquired under standardized conditions [63]. Thus, stable isotope or chemical labeling is frequently applied for quantification in proteomic methods (reviewed in [63]). While software and algorithms for protein identification are well established, quantitative analysis remains more challenging ([8, 63], see [67] for a recent survey on proteomics repositories and databases).

To date, only a restricted number of plant cell types have been profiled in a cell type-specific manner by proteomics, including guard cells, mesophyll cells, trichomes, root hair cells, leave epidermal cells, lily and rice sperm cells, different stages of pollen development in tobacco and tomato, and rice egg cells ([20, 68–70], and reviewed by [28, 35]). As compared to transcriptomics approaches, a larger amount of starting material is required. For example, approximately 40 μg of protein were isolated to study the proteome during tobacco pollen development [70]. In addition, the amount of proteins detected is typically in the range of 10-30% of the transcripts identified from the same cell or tissue type, as exemplified by a study on *Arabidopsis* pollen, in which 3'599 proteins as compared to 11'150 expressed genes were reported [71]. This quantitative difference largely reflects the difference in the sensitivity of the methods and likely only to a smaller extent meaningful biological differences. This is consistent with the identification of 13'039 proteins in a genome scale proteomics study in *Arabidopsis*, reflecting about half of all gene models [72]. Nevertheless, as only a few proteins have been identified in a previous study, e.g. from maize egg cells, these data already reflect a great improvement [73] and the rapid development starting from the shaping of the term proteomics in 1997 [74].

Protein-Protein Interaction

For studies of protein-protein interactions major methods are yeast two hybrid (Y2H), affinity purification mass spectrometry (AP-MS), or bimolecular fluorescence complementation (BiFC) (reviewed in [75]). Y2H takes advantage of the bipartite structure of the yeast GAL4 promoter consisting of two functional domains, a transcription activation domain and a DNA binding domain. In Y2H, the bait and the target protein are fused to the two functional domains, together reconstituting the functional GAL4 protein that binds to the UAS promoter to activate down-stream gene expression. Apart from a high false positive rate, the use of yeast itself is a major drawback of the method. While cell type-specific cDNA libraries could be used to profile pairwise protein interactions, the system does not truly reflect the *in vivo* state of a specific plant cell (e.g., cofactors of an interaction may be missing). Several systems similar to Y2H have been established to specifically study membrane proteins (e.g., split-ubiquitin system; [76, 77]). For AP-MS a bait protein is fused to an affinity tag for expression *in vivo*. The tagged protein of interest is subsequently purified as a complex with interacting proteins or other molecules and assayed by MS. This method is as well associated with a relatively high false positive

rate due to protein contaminants. While the method is well suitable for cell type-specific studies if the expression of the tagged protein is driven by a cell type-specific promoter, true "omics" scale profiling can hardly be achieved, as a precondition would be the cell type-specific tagging of all proteins represented in a cell. This also holds true for BiFC, where a fluorescent protein (YFP, RFP or GFP) is split in two non-fluorescent halves that are reconstructed to a fluorescent protein upon interaction of the bait and target proteins they are fused to (reviewed by [75]). While BiFC has the advantage that special and temporal interactions can be resolved, it is also associated with a high false positive rate. Consequently, methods for true cell type-specific large-scale protein-protein interaction studies in plants are lacking to date. Nonetheless, the currently available data on protein-protein interaction, as for example the recently established membrane protein interactome [78], may help resolving certain dependencies within regulatory networks (see [8] for a summary of the available databases).

Protein-DNA Interaction

Interaction between proteins and DNA comprises several functional aspects, for example histone occupancy, specific histone modifications, or transcription factor binding. These interactions may be studied using either chromatin immunoprecipitation (ChIP), or DNA adenine methyltransferase identification (Dam-ID). In both cases, the interaction of one protein (variant) with the DNA is monitored genome-wide. During the ChIP procedure, the DNA is cross-linked by formaldehyde to bound proteins before fragmentation by sonication. Chromatin fragments are then isolated with antibodies against the protein (variant) of interest. After recovery of the co-purified DNA by reverting the cross-links, the DNA sequence can be identified using microarray hybridization or high-throughput sequencing [79]. Protocols facilitating cell type-specific ChIP (chromatin immunopurification from specific cell types ChIP; CAST-ChIP) without the need for purification of the cell type of interest or a protein-specific antibody have been developed [80]. However, these protocols rely on transgenics and specific promoters. In addition, we are not aware of a report where this method has been applied in plants or to study extremely rare cell types. For Dam-ID, the protein of interest is fused to an adenine-methyltransferase of *E. coli* (Dam, [81]). Endogenous methylation of adenine is absent in most eukaryotes. Upon expression of the fusion protein, Dam is targeted to the native binding sites of the protein fused to it. This results in a localized methylation of adenines in the GATC sequence context. The regions can then be identified using methylation-sensitive restriction enzymes and microarray hybridization or high-throughput sequencing [81, 82]. Tissue or cell type-specific expression of the fusion protein can be used to overcome the need for cell isolation and has been shown to be highly specific (targeted DamID, "TaDa", [83]). The major disadvantages of the method are the requirement for transgenics and specific promoters as well as the need for optimization of the expression level to avoid untargeted methylation and toxicity of the Dam fusion protein. Thus, both approaches are currently

quite laborious and generally only applicable to model-species. Nonetheless, especially transcription factor binding is of great value for transcriptional networks [3]. If cell type-specific data is not available, previously identified transcription factor binding motifs may still help to uncover transcriptional modules [84].

Metabolomics

Due to the high complexity of plant metabolites coming from primary and secondary metabolism, the plant metabolome is highly complex to analyze. Although by far not comprehensively elucidated to date, about 200'000 different metabolites are estimated to be represented in plants (reviewed by [8]). While a variety of analysis platforms can in principle be applied for metabolite detection, nuclear magnetic resonance (NMR) and MS are the most frequently used methods [8, 85]. High resolution mapping of metabolites has recently been achieved in *Arabidopsis thaliana* roots by combining FACS with high resolution MS [45]. In addition, glandular trichomes have been used as model systems for large-scale metabolome analyses [30]. However, the major limitation of current metabolomics is the lack of a single method allowing for comprehensive measurements in terms of qualitative detection, quantitation, and spatio-temporal resolution. This is the case, as the metabolites differ significantly in their concentration, chemical properties, and analytical behavior. Two major strategies in metabolome profiling are the use of either targeted or untargeted MS (reviewed in [85]). Targeted MS relies on previous knowledge about structures and chemical properties of the metabolites of interest and combines chromatographic separation techniques, e.g. high liquid pressure chromatography (HPLC) or gas chromatography (GC) with MS techniques. In contrast, non-targeted analyses using MS without prior chromatographic separation is used to profile metabolites without prior knowledge about their abundance or structure. This method often only allows for determination of metabolic signatures, as the characterization of a specific metabolite, for example by NMR, is highly challenging. Still a key problem is therefore the availability of reference spectra and compounds for compound identification and annotation [85]. Thus the need for comprehensive databases including relevant information on the compounds, e.g. spectra, and the requirement for integration of metabolome data with other large-scale "omics" data has been noted [4]. Current online resources include the Golm Metabolome Database (`gmd.mpimp-golm.mpg.de`) and the MASSBANK Database (`www.massbank.jp`).

An alternative method to study for example metabolites at spatial resolution without the need for prior cell isolation is MALDI-imaging mass spectrometry (MSI, reviewed in [65]). For MSI, a suitable matrix is directly applied to thin tissue sections (e.g., $10 - 20\,\mu$m). The prepared tissue sections are then rasterized with a laser-beam coupled to a high mass resolution time-of-flight (TOF) mass spectrometer (reviewed in [86]). The spot size of the laser thereby determines the resolution. Only recently, technical improvements

allowed to reach resolutions required for the analysis of single cells ($< 20\,\mu$m, reviewed in [65, 85]). MSI has rarely been used in plants for proteomics, and only few studies were imaging metabolites (reviewed in [85, 86]). Examples for metabolite imaging with MSI include the measurement of wheat grain cell-wall polysaccharides ([87], $100\,\mu$m spot size), or the lipid measurements in embryos of cotton ([88], $35\,\mu$m spot size). While MSI has a great potential for cell type-specific studies for plant systems biology, it needs to be noticed that so far in MALDI only thin surface layers of $< 1\,\mu$m are sampled [65]. However, further improvements in MSI are likely to come and adaption of these methods to plant tissues may once facilitate single-cell proteomics as well as metabolomics in a wide range of species.

Systems biology approach toward plant development

As evident from the previous examples, plant cell type-specific systems biology is most advanced in cell types which allow for relatively easy isolation of high enough amounts of specific material suitable for any type of "omics" approach. For the root hairs of soybean for example, a promising method to isolate large quantities facilitating any omic analysis has recently been described and will likely be of great use [31]. In addition, for the different cell types of the *Arabidopsis* root, FACS yields sufficient material for most "omics" approaches. An advantage of these systems is that due to the use of only one isolation method, the variability imposed by it can be held constant over all experiments. It is also cost efficient as it requires less time and resources to optimize only one method as compared to several. Due to the relatively easy sample collection and their physiological roles, roots, root hairs, and trichomes are therefore excellent models to study responses to environmental stimuli, host-pathogen/symbiont interactions, metabolic pathways, or the dynamics of cellular specification and cell-cell communication in complex tissues. However, even the root may not be the an optimal model to address fundamental questions of developmental systems biology. Its main disadvantages are the long developmental time span starting very early during embryogenesis and the complex interplay within and between the different cell types of the roots but also with the above-ground tissues, and biotic and abiotic environmental factors. In contrast, a developmental model system should allow to experimentally cover the entire life-span of the whole organism. It should furthermore be rather short-lived and comprise only a limited number of developmental stages and specialized cell and tissue types to reduce complexity and increase the affordability of comprehensive studies. For comparative analyses and evolutionary systems biology approaches it would be further advantageous if the phylogeny of the model system included a broad range of organisms with gradual phenotypic changes, or with gain, loss, and alternative usage of modular building blocks. Finally, a model system is most beneficial if its understanding can lead to direct applications in, for example, production of food and pharmaceuticals.

An intuitive model for the development of an organism is the embryo. During embryogenesis, the basic body organization with an apical-basal and radial pattern is established starting from a single cell, the zygote. The mature embryo for example already contains the progenitors of the main organizers of plant growth, the primary shoot and root apical meristems (SAM and RAM), vasculature, and cotyledons (reviewed in [89]). However, it is thus already a relatively complex system composed of multiple cell and tissue types. Additional complexity is imposed by the different stages of embryo development spanning the time between the one-cellular zygote and the mature embryo. An in-depth description of embryogensis would therefore require sampling of a large variety of cell types at many time points. Nevertheless, while most transcriptional studies published so far focussed on whole tissues or entire embryos (reviewed in [90]), only recently high-quality cell type-specific transcriptomes of the early *Arabidopsis* embryo were described [91].

Alternative models for the development of organisms which are far less complex than the embryo may be the gametophytes of flowering plants: the pollen (male) and the embryo sac (female). They are typically formed from one spore (meiotic product) and at maturity they consist of only few cells and cell types including the male and female gametes, the sperm and the egg and central cell, respectively (reviewed in [92–94]). Upon double fertilization, the egg cell and the central cell fused with one sperm each give rise to the embryo and endosperm, respectively. The latter nurtures the embryo and acts as storage organ for seed reserves in several species including cereals. It is thus the most important food source for humans.

Given the sheer amount of pollen produced by a single plant, and the relatively simple isolation procedures for some of the specific cell types during pollen development, multiple cell type-specific transcriptome data sets are available from different species, including *Arabidopsis thaliana*, *Oryza sativa* (rice), *Zea mays* (maize), *Lilium longiflorum* (lily), and *Plumbago zeylanica* (doctorbush) (Table 1; reviewed in [34, 95, 96]) and several cell type-specific proteomes have recently been described from tobacco, *Lilium davidii* var. unicolor (Lanzhou lily), and tomato (Table 1; [68–70, 97]). Due to its characteristic tip-growth the pollen can also be an excellent model to study cell elongation and mechanical properties of the cell wall [98]. However, pollen development is remarkably uniform in angiosperms [99] and inter-species comparisons would therefore likely be more fruitful in gymnosperms, which show a remarkable variation in terms of the number of cell divisions between meiosis and the subsequent specification of the sperm cells [100]. In contrast to pollen, the female gametopyhtes (embryo sacs) are extremely rare and deeply embedded in the maternal floral tissue (e.g., in *Arabidopsis thaliana*, each flower contains around 50 ovules, each of which harbors only one embryo sac). Nonetheless, several cell type-specific transcriptomes (Table 2; reviewed in [34, 35], and more recent data in [53, 58, 95]) as well as a proteome analysis for rice egg cells (Table 2; [68]) are currently available. Even though more difficult to collect than the pollen, the embryo sac has certain developmental

features rendering it a highly interesting model system for plant development: (i) high evolutionary diversity within angiosperms, (ii) syncytial development (i.e., the formation of a multinucleate cell), and (iii) a process in which plants can produce asexually via seeds (gametophytic apomixis).

The mature embryo sacs of angiosperms mostly contain at least three distict cell types: the synergids required for pollen tube attraction and reception, and the two gametes, the egg and the central cell. An exception are, for example, the *Podostemaceae*, where the central cell seems to degenerate before pollen tube arrival resulting in a single fertilization event [101]. Additional antipodal cells are frequently present, but little is known about their function. In *Arabidopsis thaliana*, it has been hypothesized that they might be involved in nutrient transfer from the surrounding tissues to the embryo sac [102]. Despite the high functional similarity of the mature embryo sacs, their formation is highly diverse across different plant taxa (Figure 2, [99,103,104]). The development of the embryo sac can be devided into two steps: megasporogenesis and megagametogenesis. Megasporogenesis comprises the formation and maturation of the initial meiotic products (megaspores) from a single selected sporophytic cell, the megaspore mother cell (MMC). Megagametogenesis describes the following mitotic divisions, cellularization, and maturation of the female gametophyte. Both processes exhibit high diversity within angiosperms. Depending on the number of spores surviving and participating in megagametogenesis, megasporogenesis can be divided into *monosporic* (one spore), *bisporic* (two spores), and *tetrasporic* (all four spores). Further variation includes the location of the degenerating spores and the positioning of the spores in the *tetrasporic* types. Likewise, megagametogenesis can vary in the number of mitotic divisions, the arrangement of the nuclei/cells, and late divisions of individual cells after cellularization (e.g., *Amborella*, [105]). Comparative analysis of the structure of a wide range of embryo sacs and reconstruction of the ancestral state suggest that the embryo sacs of early angiosperms contained only four cells: two synergids, one egg cell and one central cell. It has been hypothesized that duplication of this four-celled module facilitated the emergence of the bi-nucleate central cell, which following fertilization forms an endosperm with a maternal:paternal contribution ratio of 2:1 [104–106]. This unequal parental contribution to the endosperm has received a lot of attention over the last century. As a tissue protecting and nourishing the embryo, the endosperm may be subject to adaptive processes and parental conflicts [107, 108].

An interesting aspect of megagametogenesis (and *tetrasporic* megasporogenesis) is the formation of a syncytium during the divisions of the nuclei prior to cellularization. In angiosperms, gametogenesis and early stages of endosperm development are the two major examples for the formation of a syncytium. In contrast, the plasmodial tapetum, for example, is formed by degeneration of the cell walls and the fusion of the resulting protoplasts, [109]). Unlike regular cell divisions, where the positions are relatively fixed due to the rigid cell wall, a syncytium allows for nuclear migration and for differentiation

according to positional information. Indeed, determination of cell fate in the embryo sac of *Arabidopsis thaliana* depends on the position of the nuclei as for example indicated by the *Arabidopsis retinoblastoma-related 1 (rbr1)* mutant, which produces supernumerary nuclei differentiating according to their position within the FG [110]. However, the nature of such information is still under debate. Appealing candidates may be gradients of plant hormones, such as cytokinin or auxin. For both, a role in establishing polarity during embryo sac development has been proposed (reviewed in [94]). However, an alternative or complementary hypothesis can be formulated using the analogy to the syncytial embryogenesis in *Drosophila*, where around 70% of the genes expressed during early embryogenesis show specific subcellular localization of their mRNA in the syncytium. Interestingly, specific subcellular mRNA localization peaks around the transition from syncytial to cellular development potentially reflecting the high demand for localization events [111]. Thus, a fascinating possibility is that specific subcellular localization of mRNA in the syncytial stage of the developing embryo sac may play a role in determining cell fate. A possibility to test this hypothesis would be to separately isolate specific subcellular regions (e.g., the two opposing poles) of the developing syncytial female gametophyte and to compare the transcriptional profiles of these regions with each other.

Another interesting variation of reproductive development is gametophytic apomixis. It refers to the process of asexual reproduction through seeds in the absence of fertilization (reviewed in [112]). Apomixis occurs in more than 400 plant species from around 40 genera and is likely of polyphyletic origin [113, 114]. Gametophytic apomixis involves the omission or abortion of meiosis (apomeiosis) and the formation of an embryo from an unfertilized egg (parthenogenesis), while the endosperm can be formed by autonomous development of the central cell or dependent on fertilization (pseudogamy). Depending on the mechanism of the formation of the unreduced megaspore, the resulting offspring can be genetically completely identical to the mother plant without any chromosomal rearrangements. It is thereby possible to fix complex genotypes over multiple generations without a loss in heterozygosity. While gametophytic apomixis is absent in major crop plants, engineered apomictic crops would promise great potential and economical value for plant breeding and agriculture [115–117]. From a developmental perspective, apomixis can be seen as an alteration of the sexual pathway, where certain processes are initiated too early [118]. Detailed understanding of the molecular processes and pathways governing gametogenesis during sexual and apomictic reproduction is therefore a precondition to engineer apomixis in crop plants. In evolutionary terms, apomixis is a highly interesting trait. On one hand, it allows for dispersial through seeds without the need for a sexual partner [119] and may therefore be advantageous for colonization of new habitats [120]. On the other hand, the trade-off for this clonal reproduction appears to be very costly. Apomicts may accumulate deleterious mutations over many generations [121] and their populations are likely of low genetic variability, which reduces their potential to adapt to a changing environment.

Taken together the natural variation in sexual and apomictic species, the female gametophyte of angiosperms can be seen as an excellent model system to study fundamental developmental processes and evolutionary aspects of plant development and biology with high importance for agriculture. Its simple organization and the relatively few developmental stages would allow for an in-depth analysis of various species enabling evolutionary comparisons at the whole genome level. Given the high diversity, inter-species comparisons may identify genes and genetic networks involved in the emergence of evolutionary novelties such as the unequal genetic contribution of the two parents to the endosperm, or gametophytic apomixis. Deciphering the evolutionary mechanism underlying these processes may also provide an answer to the long standing question on how useful research on model organisms is for crop improvement. However, the small size and inaccessibility of the cell types of the developing and mature embryo sacs make the isolation and subsequent application of omics methods very difficult. Aside the challenges associated with data integration and analysis, data generation is hence a major limiting factor. In general, the main obstacle with most approaches is the number of cells required for in-depth profiling of a certain molecule (e.g., protein or metabolite). This may be overcome by either increased sensitivity of the profiling method, or through simplified collection of a large amount of cells. However, most high-throughput isolation methods (e.g., for FACS/FANS/INTACT) rely on the existence of a specific marker (i.e., a cell type-specific promoter) and the possibility to generate transgenic plants. In addition, typically a certain abundance of the cell type of interest in the sample is required for efficient sorting and purification. Given that these preconditions are generally missing in the case of low abundant cell types of of non-model organisms, it is likely the increase in sensitivity and the development of novel profiling methods from which plant systems biology will profit most. In the following sections, we will therefore focus on a subset of "omics" approaches, which are readily available or which bear great future potential for routine large-scale *in vivo* profiling of specific cell types of the female gametophytes of flowering plants. The examples given are restricted to studies on specific cell types of the female gametophytes of angiosperms.

Transcriptome

Transcriptomics is clearly the most frequently used and currently the most robust "omics" approach to study female gametophyte and plant reproductive development. Following the early transcriptional profiling with low-throughput technologies (early EST sequencing projects, reviewed in [35]), cell type-specific transcriptomes were generated for all cell types of the mature female gametophyte and the megaspore mother cell (MMC) of *Arabidopsis thaliana* [55–57], the egg cell and the synergids for rice [95, 122], all cell types of the mature embryo sac and the apomictic initial cell (AIC) of *Boechera gunnisoniana* (a close apomictic relative of *Arabidopsis thaliana* where an AIC is specified instead of a sexual MMC, [58]), and the aposporous initial cell (AI) of *Hieracium praealtum* (Hawk-

weed, where the apomictic embryo sac is formed by an additional sporophytic cell (AI) developing adjacent to the sexual reproductive lineage, [53]; Table 2). Given the requirement to establish a specific gene expression profile for cell differentiation and specification, transcriptomics is also especially suitable as a first approach toward an unknown species because it provides a comprehensive snapshot of the cellular instruction machinery. It further enables the identification of cell type-specific markers and can thus provide a basis for other approaches like detailed molecular and mechanistic studies. The advantage of transcriptional profiling as compared to proteomic studies is the possibility to amplify the material prior to detection. Several RNA-Seq protocols allow transcriptional profiling of singe cells corresponding to as little as about 10 pg of total RNA (reviewed in [123]) . The low detection limit facilitates the use of relatively low throughput isolation methods, such as laser-assisted microdissection or manual microdissection allowing for profiling of specific cell types of embryo sacs of model and non-model species [35, 53, 58]. A current drawback of the amplification strategy is the introduction of potential quantification biases. A possible solution may be unique molecular identifiers (UMI). These are short sequences with random nucleotides (e.g., 1'024 different UMIs with 5 random nucleotides), which are used to label initial cDNA molecules prior to amplification. An excess of UMIs compared to the number of identical cDNAs ensures that each combination of a given UMI with a certain cDNA is unique. After amplification and sequencing, this can be used to differentiate between individual molecules in the initial cDNA pool and duplicates originating from cDNA amplification (i.e., to count molecules instead of reads, [124]). An interesting approach for future studies may be fluorescent *in situ* RNA sequencing (FISSEQ), in which stably cross-linked cDNA amplicons are sequenced directly within a biological sample, thereby not only quantifying gene expression, but also detecting the subcellular localization of the transcripts [125]. Improvement of the method and the adaption of the method for plant tissues would thus undoubtfully be a major advance in cell type-specific transcriptional profiling.

Proteome and Metabolome

Proteomics and metabolomics on specific cell types is substantially more challenging than transcriptomics. A current limitation for cell type-specific proteomics is the frequently large discrepancy between the number of detected proteins compared to the number of expressed genes due to the low sensitivity of proteomics methods toward low-abundant proteins. Additional complexity arises by the presence of a wide range of post-translational modifications like phosphorylation or glycosylation. Aside two early examples identifying only the major proteins in the egg cells of maize and rice (6 and 4 proteins, [73, 126]), we are only aware of the recent description of the egg cell proteome in rice, where 2'138 proteins were identified using around 500 egg cells ([68]; Table 2). In the same study, Similarly, 2'179 proteins were identified starting from 30'000 isolated sperm cells (Table 1; [68]). Given the further improvements of the sensitivity of mass spectrometers, the

example demonstrates that proteomics of purified cells of the female gametopyhte should already be possible for cases where enough material can be collected. Mechanical or manual isolation of female gametes was reported for a variety of species including barley, wheat, rape, maize, tabacco, *Alstroemeria*, and *Arabidopsis thaliana* [127–133]. In most of these species, we anticipate that the protocols would already allow for isolation of sufficient material for mass spectrometry-based proteomics. Another promising approach for future experiments may be MSI, circumventing the need of (laborious) cell purification.

Methylome

DNA cytosine methylation (5mC) plays an important role in the epigenetic regulation of plant genomes. The current method of choice for methylome profiling is whole-genome bisulfite sequencing (WGBS). In brief, DNA is incubated with bisulfite converting all unmethylated cytosines to uracils, which are then identified as thymines during sequencing (reviewed in [134]) . Compared to the profiling of other epigenetic marks, such as histone modifications, WGBS has two major advantages. It does not require the use of transgenics or antibodies and recently developed methods facilitate WGBS on as little as 125 pg of DNA (post-bisulfite adaptor tagging (PBAT), [135]; 20 pg diluted *Arabidopsis thaliana* DNA with modified protocol, in-house data). Even for a plant with a small genome, for example 130 Mb in *Arabidopsis thaliana*, this corresponds only to around 900 haploid nuclei. While WGBS has not yet been reported for isolated cells of the female gametophyte, bisulfite sequencing of specific sequences has already been applied for *Arabidopsis thaliana* central cells and synergids isolated by LAM [136, 137]. It would likely be possible to combine LAM or manual microdissection with WBGS. This would thus allow for methylome profiling of model as well as non-model species. Importantly, this might allow to gain novel insights into the molecular basis underlying heterosis, characterized by superior characteristics of F1 hybrid plants as compared to their parents. While epigenetic regulatory pathways are likely important for heterosis, their precise involvement remains elusive to date [138]. Understanding of the regulatory mechanisms governing heterosis is of great interest for plant breeding and crop production. Importantly, gametophytic development and early stages of embryogenesis are likely important for establishing heterosis.

Conclusion and Perspectives

To date, cell type-specific systems biology in plant sciences is frequently constrained by the difficulties associated with the isolation of the cell type of interest in large enough amounts. Robust and simple isolation methods exist for only few cell types. Consequently, the comprehensive profiling of all cell types of an organism with different large-scale profiling methods allowing the detailed understanding of all biological processes ongoing in the biological system is still an unreached goal. While the in-depth understanding of complex

organism over their livespan is a major aim for systems biology, the use of simple model organisms bears advantages, given the persisting technical limitations. We introduce the female gametophyte of angiosperms as an attractive model system for future systems biology approaches in plant development. Apart from its relatively simple organisation, it is of great biological and agronomical importance, for example in respect to crop seed production and plant breeding. However, most high-throughput isolation methods with broader application (e.g., FACS/FANS/INTACT) are currently limited to model organisms (e.g., *Arabidopsis thaliana*). However, a biological system may be best understood in the context of evolution. In addition, detailed understanding of cellular processes in all major agriculturally important species including wheat, where an additional challenge is the genome size and the hexaploid nature, are a precondition for targeted crop improvement. Such studies would thus not only be of potential applied value, but also would help to understand the common concepts and divergent mechanisms active in different species. Therefore, methods facilitating large-scale profiling of specific cell types in model as well as non-model organism are critical. Parallel high-throughput profiling of several organisms covering a phenotypic gradient, or including gain, loss, and alternative usage of modular building blocks along the phylogeny may then enable evolutionary systems biology. This may ultimately help to reconstruct the emergence of evolutionary novelties and to find the underlying genetic and molecular networks. Such an understanding would in turn allow the control of the underlying processes with an unprecedented resolution. In perspective, this can be an important precondition for targeted improvement of crop species, including the engineering of apomixis into crop plants.

Even though the isolation of the individual cell types is currently still very challenging, the dramatic technical advances observed over the past few years in, for example, transcriptional profiling are clear indications for the rapid development and improvement of the large-scale profiling technologies. In this light, we emphasized methods for transcriptomics, proteomics, metabolomics, and methylomics, in which we see great future potential. It is important to note that novel methods allowing large-scale profiling without prior cell isolation, for example MSI or FISSEQ, are very promising for future applications. The growing amount of data and data types also points out the need for computational solutions addressing the problems of data storage, integration, and analysis (see [3–6, 36–40]). The current situation, in which data sometimes remain unpublished, are frequently poorly annotated, and widely dispersed on specialized databases, may be taken as motivation to develop integrative computational platforms specifically focussing on future data. Considering the almost exponential growth of biological data over the last years [1,5], these platforms may also ignore data from the past to allow for innovative solutions. In this context, standardized data formats and annotation, easily accessible databases, powerful data mining tools, user-friendly and freely available software, as well as scalable storage platforms are the current and future demands in systems biology [5,36].

Disclosure/Conflict-of-Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

Funding: Work on gametophyte development, apomixis and epigenetic gene regulation in UG's laboratory is supported by the University of Zürich, and by grants from the "Staatssekretariat für Bildung und Forschung" in the framework of COST action FA0903 (to UG and AS), the Swiss National Science Foundation (to UG) and the European Research Council (to UG).

References

[1] Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: Systems biology. Annual Review of Genomics and Human Genetics 2: 343–372.

[2] Kitano H (2002) Systems biology: A brief overview. Science 295: 1662–1664.

[3] Yuan JS, Galbraith DW, Dai SY, Griffin P, Stewart CN (2008) Plant systems biology comes of age. Trends in Plant Science 13: 165–171.

[4] Fukushima A, Kusano M, Redestig H, Arita M, Saito K (2009) Integrated omics approaches in plant systems biology. Current Opinion in Chemical Biology 13: 532–538.

[5] Chuang HY, Hofree M, Ideker T (2010) A decade of systems biology. Annual Review of Cell and Developmental Biology 26: 721–44.

[6] Katari MS, Nowicki SD, Aceituno FF, Nero D, Kelfer J, et al. (2010) Virtual Plant: A software platform to support systems biology research. Plant Physiology 152: 500–515.

[7] Weckwerth W (2011) Green systems biology – from single genomes, proteomes and metabolomes to ecosystems research and biotechnology. Journal of Proteomics 75: 284–305.

[8] Sheth BP, Thaker VS (2014) Plant systems biology: insights, advandes and challenges. Planta 240: 33–54.

[9] Pelkmans L (2012) Using cell-to-cell variability – a new era in molecular biology. Science 336: 425–426.

[10] Brown JA, Sherlock G, Myers CL, Burrows NM, Deng C, et al. (2006) Global analysis of gene function in yeast by quantitative phenotypic profiling. Molecular Systems Biology 2: 2006 0001.

[11] Süel GM, Kulkarni RP, Dworkin J, Garcia-Ojalvo J, Elowitz MB (2007) Tunability and noise depence in differentiation dynamics. Science 315: 1716–1719.

[12] Konstantinidis KT, Serres MH, Romine MF, Rodrigues JLM, Auchtung J, et al. (2009) Comparative systems biology across an evolutionary gradient within the *Shewanella* genus. PNAS 106: 15909–15914.

[13] Soyer OS (2012) Evolutionary Systems Biology, volume 751 of *Advances in Experimental Medicine and Biology*. Heidelberg, Ger: Springer.

[14] Boone C (2014) Yeast systems biology: Our best shot at modeling a cell. Genetics 198: 435–437.

[15] Bostein D, Fink GR (2011) Yeast: An experimental organism for 21st century biology. Genetics 189: 695–704.

[16] Österlund T, Nookaew I, Nielsen J (2012) Fifteen years of large scale metabolic modeling of yeast: Developments and impacts. Biotechnology Advances 30: 979–988.

[17] Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, et al. (2010) The genetic landscape of a cell. Science 327: 425–431.

[18] Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296: 752–755.

[19] Ishida T, Kurata T, Okada K, Wada T (2008) A genetic regulatory network in the development of trichomes and root hairs. Annual Review of Plant Biology 59: 365–386.

[20] Brechenmacher L, Lee J, Sachdev S, Song Z, Nguyen THN, et al. (2009) Establishment of a protein reference map for soybean root hair cells. Plant Physiology 149: 670–682.

[21] Brechenmacher L, Lei Z, Libault M, Findley S, Sugawara M, et al. (2010) Soybean metabolites regulated in root hairs in response to the symbiotic bacterium *Bradyrhizobium japonicum*. Plant Physiology 153: 1808–1822.

[22] Dai X, Wang G, Yang DS, Tang Y, Broun P, et al. (2010) TrichOME: a comparative omics database for plant trichomes. Plant Physiology 152: 44–54.

[23] Libault M, Brechenmacher L, Cheng J, Xu D, Stacey G (2010) Root hair systems biology. Trends in Plant Science 15: 641–650.

[24] Libault M, Farmer A, Brechenmacher L, Drnevich J, Langley RJ, et al. (2010) Complete transcriptome of the soybean root hair cell, a single-cell model, and its alteration in response to *Bradyrhizobium japonicum* infection. Plant Physiology 152: 541–552.

[25] Schilmiller AL, Miner DP, Larson M, McDowell E, Gang DR, et al. (2010) Studies of a biochemical factory: tomato trichome deep expressed sequence tag sequencing and proteomics. Plant Physiology 153: 1212–1223.

[26] Nestler J, Schütz W, Hochholdinger F (2011) Conserved and unique features of the maize (*Zea mays* L) root hair proteome. Journal of Proteome Research 10: 2525–2537.

[27] Van Cutsem E, Simonart G, Degand H, Faber AM, Morsomme P, et al. (2011) Gel-based and gel-free proteomic analysis of *Nicotiana tabacum* trichomes identifies proteins involved in secondary metabolism and in the (a)biotic stress response. Proteomics 11: 440–454.

[28] Dai S, Chen S (2012) Single-cell-type proteomics: Toward a holistic understanding of plant function. Molecular & Cellular Proteomics 11: 1622–1630.

[29] Rogers ED, Jackson T, Moussaieff A, Aharoni A, Benfey PN (2012) Cell type-specific transcriptional profiling: implications for metabolite profiling. The Plant Journal 70: 5–17.

[30] Tissier A (2012) Glandular trichomes: what comes after expressed sequence tags? The Plant Journal 70: 51–68.

[31] Qiao Z, Libault M (2013) Unleashing the potential of the root hair cell as a single plant cell type model in root systems biology. Frontiers in Plant Science 4: 484.

[32] Kehr J (2001) High resolution spation analysis of plant systems. Current Opinion in Plant Biology 4: 197–201.

[33] Taylor-Teeples M, Ron M, Brady SM (2011) Novel biological insights revealed from cell type-specific expression profiling. Current Opinion in Plant Biology 14: 1–7.

[34] Schmidt A, Schmid MW, Grossniklaus U (2012) Analysis of plant germline development by high-throughput RNA profiling: technical advances and new insights. The Plant Journal 70: 18-29.

[35] Wuest SE, Schmid MW, Grossniklaus U (2013) Cell-specific expression profiling of rare cell types as exemplified by its impact on our understanding of female gametophyte development. Current Opinion in Plant Biology 16: 41–49.

[36] Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, et al. (2014) Data integration in the era of omics: current and future challenges. BMC Systems Biology 8: I1.

[37] Ahrens CH, Wagner U, Rehrauer HK, Türker C, Schlapbach R (2007) Current challenges and approaches for the synergistic use of systems biology data in the scientific community. Experientia Supplementum 97: 277–307.

[38] Liberman LA, Sozzani R, Benfey PN (2012) Integrative systems biology: an attempt to describe a simple weed. Current Opinion in Plant Biology 15: 162–167.

[39] Robinson SW, Fernandes M, Husi H (2014) Current advances in systems and integrative biology. Computational and Structural Biotechnology Journal 11: 35–46.

[40] Fukushima A, Kanaya S, Nishida K (2014) Integrated network analysis and effective tools in plant systems biology. Frontiers in Plant Science 5: 598.

[41] Zhu M, Dai S, McClung S, Yan X, Chen S (2009) Functional differentiation of *Brassica napus* guard cells and mesophyll cells revealed by comparative proteomics. Molecular & Cellular Proteomics 8: 752–766.

[42] Hu TX, Yu M, Zhao J (2011) Techniques of cell type-specific transcriptome analysis and applications in researches of plant sexual reproduction. Frontiers in Biology 6: 31–39.

[43] Petricka JJ, Schauer MA, Megraw M, Breakfield NW, Thompson JW, et al. (2012) The protein expression landscape of the *Arabidopsis* root. PNAS 109: 6811–6818.

[44] Brady SM, Orlando DA, Lee JY, Wang JY, Koch J, et al. (2007) A high-resolution root spatiotemporal map reveals dominant expression patterns. Science 318: 801–806.

[45] Moussaieff A, Rogachev I, Brodsky L, Malitsky S, Toal TW, et al. (2013) High-resolution metabolic mapping of cell types in plant roots. PNAS 110: E1232–E1241.

[46] Benfey PN (2012) Toward a systems analysis of the root. Cold Spring Harbor Symposia on Quantitative Biology 77: 91–96.

[47] Weinhofer I, Hehenberger E, Roszak P, Hennig L, Köhler C (2010) H3K27me3 profiling of the endosperm implies exclusion of polycomb group protein targeting by DNA methylation. PLOS Genetics 6: e1001152.

[48] Weinhofer I, Köhler C (2014) Endosperm-specific chromatin profiling by fluorescence-activated nuclei sorting and ChIP-on-chip. Methods in Molecular Biology 1112: 105–115.

[49] Evrard A, Bargmann BO, Birnbaum KD, Tester M, Baumann U, et al. (2012) Fluorescence-activated cell sorting for analysis of cell type-specific responses to salinity stress in *Arabidopsis* and rice. Methods in Molecular Biology 913: 265–276.

[50] Deal RB, Henikoff S (2011) The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana*. Nature Protocols 6: 56–68.

[51] Bailey-Serres J (2013) Microgenomics: Genome-scale, cell-specific monitoring of multiple gene regulation tiers. Annual Review of Plant Biology 64: 293–325.

[52] Tucker MR, Okada T, Hu Y, Scholefield A, Taylor JM, et al. (2012) Somatic small RNA pathways promote the mitotic events of megagametogenensis during female reproductive development in *Arabidopsis*. Development 139: 1399–1404.

[53] Okada T, Hu Y, Tucker MR, Taylor JM, Johnson SD, et al. (2013) Enlarging cells initiating apomixis in *Hieractium praealtum* transition to an embryo sac program prior to entering mitosis. Plant Physiology 163: 216–231.

[54] Wuest SE, Grossniklaus U (2014) Laser-assisted microdissection applied to floral tissues. Methods in Molecular Biology 1110: 329–344.

[55] Wuest SE, Vijverberg K, Schmidt A, Weiss M, Gheyselinck J, et al. (2010) *Arabidopsis* female gametophyte gene expression map reveals similarities between plant and animal gametes. Current Biology 20: 1–7.

[56] Schmidt A, Wuest SE, Vijverberg K, Baroux C, Kleen D, et al. (2011) Transcriptome analysis of the *Arabidopsis* megaspore mother cell uncovers the importance of RNA helicases for plant germ line development. PLOS Biology 9: e1001155.

[57] Schmid MW, Schmidt A, Klostermeier UC, Barann M, Rosenstiel P, et al. (2012) A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing. PLOS ONE 7: e29685.

[58] Schmidt A, Schmid MW, Klostermeier UC, Qi W, Guthörl D, et al. (2014) Apomictic and sexual germline development differ with respect to cell cycle, transcriptional, hormonal and epigenetic regulation. PLOS Genetics 10: e1004476.

[59] Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, et al. (2005) A gene expression map of *Arabidopsis thaliana* development. Nature Genetics 37: 501–506.

[60] Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, et al. (2003) A gene expression map of the *Arabidopsis* root. Science 302: 1956–1960.

[61] Mardis ER (2013) Next-generation sequencing platforms. Annual Review of Analytical Chemistry 6: 287–303.

[62] Schmid MW, Grossniklaus U (2015) Rcount: simple and flexible RNA-Seq read counting. Bioinformatics 31: 436–437.

[63] Schulze WX, Usadel B (2010) Quantitation in mass-spectrometry-based proteomics. Annual Review of Plant Biology 61: 491–516.

[64] Hollenbeck TP, Siuzdak G, Blackledge RD (1999) Electrospray and MALDI mass spectrometry in the identification of spermicides in criminal investigations. Journal of Forensic Sciences 44: 793–788.

[65] Lee YJ, Perdian DC, Song Z, Yeung ES, Nikolau BJ (2012) Use of mass spectrometry for imaging metabolites in plants. The Plant Journal 70: 81–95.

[66] Perry RH, G C, J NR (2008) Orbitrap mass spectrometry: Instrumentation, ion motion and applications. Mass Spectrometry Reviews 27: 661–699.

[67] Sakata K, Komatsu S (2014) Plant proteomics: from genome sequencing to proteome databases and repositories. Methods in Molecular Biology 1072: 29–42.

[68] Abiko M, Furuta K, Yamauchi Y, Fujita C, Taoka M, et al. (2013) Identification of proteins enriched in rice egg or sperm cells by single-cell proteomics. PLOS ONE 8: e69578.

[69] Chaturvedi P, Ischebeck T, Egelhofer V, Lichtscheidl I, Weckwerth W (2013) Cell-specific analysis of the tomato pollen proteome from pollen mother cell to mature pollen provides evidence for developmental priming. Journal of Proteome Research 12: 4892–4903.

[70] Ischebeck T, Valledor L, Lyon D, Gingl S, Nagler M, et al. (2014) Comprehensive cell-specific protein analysis in early and late pollen development from diploid microsporocytes to pollen tube growth. Molecular & Cellular Proteomics 13: 295–310.

[71] Grobei MA, Qeli E, Brunner E, Rehrauer H, Zhang R, et al. (2009) Deterministic protein inference for shotgun proteomics data provides new insights into *Arabidopsis* pollen development and function. Genome Research 19: 1786–1800.

[72] Baerenfaller K, Grossman J, Grobei MA, Hull R, Hirsch-Hoffmann M, et al. (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. Science 320: 938–941.

[73] Okamoto T, Higuchi K, Shinkawa T, Isobe T, Lörz H, et al. (2004) Identification of major proteins in maize egg cells. Plant Cell Physiology 45: 1406–1412.

[74] James P (1997) Protein identification in the post-genome era: the rapid rise of proteomics. Quarterly Reviews of Biophysics 30: 279–331.

[75] Zhang Y, Gao P, Yuan JS (2010) Plant protein-protein interaction network and interactome. Current Genomics 11: 40–46.

[76] Obrdlik P, El-Bakkoury M, Hamacher T, Cappellaro C, Vilarino C, et al. (2004) K^+ channel interactions detected by a genetic system optimized for systematic studies of membrane protein interactions. PNAS 101: 12242–12247.

[77] Chen C, Farmer AD, Langley RJ, Mudge J, Crow JA, et al. (2010) Meiosis-specific gene discovery in plants: RNA-Seq applied to isolated *Arabidopsis* male meiocytes. BMC Plant Biology 10: 280.

[78] Chen J, Lalonde S, Obrdlik P, Vatani AN, Pasra SA, et al. (2012) Uncovering *Arabidopsis* membrane protein interactome enriched in transporters using mating-based split ubiquitin assays and classification models. Frontiers in Plant Science 3: 124.

[79] He G, Elling AA, Deng XW (2011) The epigenome and plant development. Annual Review of Plant Biology 62: 411–435.

[80] Schauer T, Schwalie PC, Handley A, Margulies CE, Flicek P, et al. (2013) CAST-ChIP maps cell-type-specific chromatin states in the *Drosophila* central nervous system. Cell Reports 5: 271–282.

[81] Greil F, Moorman C, Van Steensel B (2006) DamID: Mapping of *in vivo* protein-genome interactions using tethered DNA adenine methyltransferase. Methods in Enzymology 410: 342–359.

[82] Luo SD, Shi GW, Baker BS (2011) Direct targets of the *D melanogaster* DSX^F protein and the evolution of sexual development. Development 138: 2761–2771.

[83] Southall TD, Gold KS, Egger B, Davidson CM, Caygill EE, et al. (2013) Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: Assaying RNA pol II occupancy in neural stem cells. Developmental Cell 26: 101–112.

[84] Diez D, Hutchins AP, Miranda-Saavedra D (2014) Systematic identification of transcriptional regulatory modules from protein-protein interaction networks. Nucleic Acids Research 42: e6.

[85] Kueger S, Steinhauser D, Willmitzer L, Giavalisco P (2012) High-resolution plant metabolomics: from mass spectral features to metabolites and from whole-cell analysis to subcellular metabolite distributions. The Plant Journal 70: 39–50.

[86] Kaspar S, Peukert M, Svatos A, Matros A, Mock HP (2011) MALDI-imaging mass spectrometry - an emerging technique in plant biology. Proteomics 11: 1840–1850.

[87] Veličković D, Ropartz D, Guillon F, Saulnier L, Rogniaux H (2014) New insights into the structural and spatial variability of cell-wall polysaccharides during wheat grain development, as revealed through MALDI mass spectrometry imaging. Journal of Experimental Botany 65: 2079–2091.

[88] Horn PJ, Korte AR, Neogi PB, Love E, Fuchs J, et al. (2012) Spatial mapping of lipids at cellular resolution in embryos of cotton. The Plant Cell 24: 622–636.

[89] Lau S, Slane D, Herud O, Kong J, Jürgens G (2012) Early embryogenesis in flowering plants: Setting up the basic body pattern. Annual Review of Plant Biology 63: 6 1–6 24.

[90] Palovaara J, Saiga S, Weijers D (2013) Transcriptomics approaches in the early *Arabidopsis* embryo. Trends in Plant Science 18: 514–521.

[91] Slane D, Kong J, Berendzen KW, Kilian J, Henschen A, et al. (2014) Cell type-specific transcriptome analysis in the early *Arabidopsis thaliana* embryo. Development 141: 4831–4840.

[92] Yang WC, Shi DQ, Chen YH (2010) Female gametophyte development in flowering plants. Annual Review of Plant Biology 61: 89–108.

[93] Twell D (2011) Male gametogenesis and germline specification in flowering plants. Sexual Plant Reproduction 24: 149–160.

[94] Schmidt A, Schmid MW, Grossniklaus U (2015) Plant germline formation: molecular insights define common concepts and illustrate developmental flexibility in apomictic and sexual reproduction. Development 142: 229–241.

[95] Anderson SN, Johnson CS, Jones DS, Conrad LJ, Gou X, et al. (2013) Transcriptomes of isolated *Oryza sativa* gametes characterized by deep sequencing: evidence for distinct sex-dependent chromatin and epigenetic states before fertilization. The Plant Journal 76: 729–741.

[96] Dukowic-Schulze S, Sundararajan A, Mudge J, Ramaraj T, Farmer AD, et al. (2014) The transcriptome landscape of early maize meiosis. BMC Plant Biology 14: 118.

[97] Zhao X, Yang N, Wang T (2013) Comparative proteomic analysis of generative and sperm cells reveals molecular characteristics associated with sperm development and function specialization. Journal of Proteome Research 12: 5058–5071.

[98] Vogler H, Draeger C, Weber A, Felekis D, Eichenberger C, et al. (2013) The pollen tube: a soft shell with a hard core. The Plant Journal 73: 617–627.

[99] Maheshwari P (1950) An Introduction to the Embryology of Angiosperms. New York, USA: McGraw-Hill.

[100] Fernando DD, Quinn CR, Brenner ED, Owens JN (2010) Male gametophyte development and evolution in extant gymnosperms. International Journal of Plant Developmental Biology 4 (Special Issue 1): 47–63.

[101] Sehgal A, Mann N, Mohan Ram HY (2014) Structural and developmental variability in the female gametophyte of *Griffithella hookeriana*, *Polypleurum stylosum*, and *Zeylanidium lichenoides* and its bearing on the occurrence of single fertilization in Podostemaceae. Plant Reproduction 27: 205–223.

[102] Raghavan V (1997) Molecular Embryology of Flowering Plants. Cambridge, UK: Cambridge University Press.

[103] Huang BQ, Russell SD (1992) Female germ unit: Organization, isolation, and function. International Review of Cytology 140: 233–293.

[104] Williams JH, Friedman WE (2004) The four-celled female gametophyte of *Illicium* (Illiciaceae; Austrobaileyales): implications for understanding the origin and early evolution of monocots, eumagnoliids, and eudicots. American Journal of Botany 91: 332–351.

[105] Friedman WE (2006) Embryological evidence for developmental lability during early angiosperm evolution. Molecular Systems Biology 2: 2006 0001.

[106] Friedman WE, Ryerson KC (2009) Reconstructing the ancestral female gametophyte of angiosperms: insights from *Amborella* and other ancient lineages of flowering plants. American Journal of Botany 96: 129–143.

[107] Haig D, Westoby M (1989) Parent-specific gene expression and the triploid endosperm. The American Naturalist 134: 147–155.

[108] Baroux C, Spillane C, Grossniklaus U (2002) Evolutionary origins of the endosperm in flowering plants. Genome Biology 3: reviews1026 1-reviews1026 5.

[109] Furness CA, Rudall PJ (1998) The tapetum and systematics in monocotyledons. Botanical Review 64: 201–239.

[110] Ebel C, Mariconti L, Gruissem W (2004) Plant retinoblastoma homologues control nuclear proliferation in the female gametophyte. Nature 429: 776–780.

[111] Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, et al. (2007) Global analyses of mRNA localization reveals a prominent role in organizing cellular architecture and function. Cell 131: 174–187.

[112] Koltunow AM, Grossniklaus U (2003) Apomixis: a developmental perspective. Annual Review of Plant Biology 54: 547–574.

[113] Asker SE, Jerling L (1992) Apomixis in Plants. London, UK: CRC Press.

[114] Carman JG (1997) Asynchronous expression of duplicate genes in angiosperms may cause apomixis, bispory, tetraspory, and polyembryony. Biological Journal of The Linnean Society 61: 51–94.

[115] Koltunow AM, Bicknell RA, Chaudhury AM (1995) Apomixis: Molecular strategies for the generation of genetically identical seeds without fertilization. Plant Physiology 108: 1345–1352.

[116] Vielle-Calzada JP, Crane C, Stelly DM (1996) Apomixis – the asexual revolution. Science 274: 1322–1323.

[117] Grossniklaus U, Koltunow A, van Lookeren Campagne M (1998) A bright future for apomixis. Trends in Plant Sciences 3: 415–416.

[118] Koltunow AM (1993) Apomixis: embryo sacs and embryos formed without meiosis or fertilization in ovules. The Plant Cell 5: 1425–1437.

[119] Smith JM (1978) The Evolution of Sex. Cambridge, UK: Cambridge University Press.

[120] Tomlinson J (1966) The advantages of hermaphroditism and parthenogenesis. Journal of Theoretical Biology 11: 54–58.

[121] Muller HJ (1964) The relation of recombination of mutational advance. Mutation Research 106: 2–9.

[122] Ohnishi T, Takanashi H, Mogi M, Takahashi H, Kikuchi S, et al. (2011) Distinct gene expression profiles in egg and synergid cells of rice as revealed by cell type-specific microarrays. Plant Physiology 155: 881–891.

[123] Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, et al. (2014) Library construction for next-generation sequencing: Overviews and challenges. Biotechniques 56: 61–77.

[124] Islam S, Zeisel A, Joost S, La Manno G, Zajac P, et al. (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods 11: 163–166.

[125] Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, et al. (2014) Highly multiplexed subcellular RNA sequencing in situ. Science 343: 1360–1363.

[126] Uchiumi T, Shinkawa T, Isobe T, Okamoto T (2007) Identification of the major protein components of rice egg cells. Journal of Plant Research 120: 575–579.

[127] Holm PB, Knudsen S, Mouritzen P, Negri D, Olsen FL, et al. (1994) Regeneration of fertile barley plants from mechanically isolated protoplasts of the fertilized egg cell. The Plant Cell 6: 531–543.

[128] Kovács M, Barnabás B, Kranz E (1994) The isolation of viable egg cells of wheat (*Triticum aestivum* l). Sexual Plant Reproduction 7: 311–312.

[129] Katoh N, Lörz H, Kranz E (1997) Isolation of viable egg cells of rape (*Brassica napus* l). Zygote 5: 31–33.

[130] Kranz E, Bautor J, Lörz H (1991) *In vitro* fertilization of single, isolated gametes of maize mediated by electrofusion. Sexual Plant Reproduction 4: 12–16.

[131] Tian HQ, Russell SD (1997) Micromanipulation of male and female gametes of *Nicotiana tabacum*: I isolation of gametes. Plant Cell Reports 16: 555–560.

[132] Hoshino Y, Murata N, Shinoda K (2006) Isolation of individual egg cells and zygotes in *Alstroemeria* followed by manual selection with a microcapillary-connected micropump. Annals of Botany 97: 1139–1144.

[133] Jullien PE, Susaki D, Yelagandula R, Higashiyama T, Berger F (2012) DNA methylation dynamics during sexual reproduction in *Arabidopsis thaliana*. Current Biology 22: 1825–1830.

[134] Ji L, Neumann DA, Schmitz RJ (2015) Crop epigenomics: Identifying, unlocking, and harnessing cryptic variation in crop genomes. Molecular Plant in press.

[135] Miura F, Enomoto Y, Dairiki R, Ito T (2012) Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. Nucleic Acids Research 40: e136.

[136] Wöhrmann HJP, Gagliardini V, Raissig MT, Wehrle W, Arand J, et al. (2012) Identification of a DNA methylation-independent imprinting control region at the *Arabidopsis MEDEA* locus. Genes & Development 26: 1837–1850.

[137] You W, Tyczewska A, Spencer M, Daxinger L, Schmid MW, et al. (2012) Atypical DNA methylation of genes encoding cysteine-rich peptides in *Arabidopsis thaliana*. BMC Plant Biology 12: 51.

[138] Chen ZJ (2013) Genomic and epigenetic insights into the molecular bases of heterosis. Nature Reviews Genetics 14: 471–482.

[139] Tang X, Zhang ZY, Zhang WJ, Zhao XM, Li X, et al. (2010) Global gene profiling of laser-captured pollen mother cells indicates molecular pathways and gene subfamilies involved in rice meiosis. Plant Physiology 154: 1855–1870.

[140] Yang H, Lu P, Wang Y, Ma H (2011) The transcriptome landscape of *Arabidopsis* male meiocytes from high-throughput sequencing: the complexity and evolution of the meiotic process. The Plant Journal 65: 503–516.

[141] Libeau P, Durandet M, Granier F, Marquis C, Berthomé R, et al. (2011) Gene expression profiling of *Arabidopsis* meiocytes. Plant Biology 13: 784–793.

[142] Honys D, Twell D (2004) Transcriptome analysis of haploid male gametophyte development in *Arabidopsis*. Genome Biology 5: R85.

[143] Wei LQ, Xu WY, Deng ZY, Su Z, Xue Y, et al. (2010) Genome-scale analysis and comparison of gene expression profiles in developing and germinated pollen in *Oryza sativa*. BMC Genomics 11: 338.

[144] Okada T, Singh MB, Bhalla PL (2007) Transcriptome profiling of *Lilium longiflorum* generative cells by cDNA microarray. Plant Cell Reports 26: 1045–1052.

[145] Borges F, Gomes G, Gardner R, Moreno N, McCormick S, et al. (2008) Comparative transcriptomics of *Arabidopsis* sperm cells. Plant Physiology 148: 1168–1181.

[146] Gou X, Yuan T, Wei X, Russell SD (2009) Gene expression in the dimorphic sperm cells of *Plumbago zeylanica*: transcript profiling, diversity, and relationship to cell type. The Plant Journal 60: 33–47.

Figures

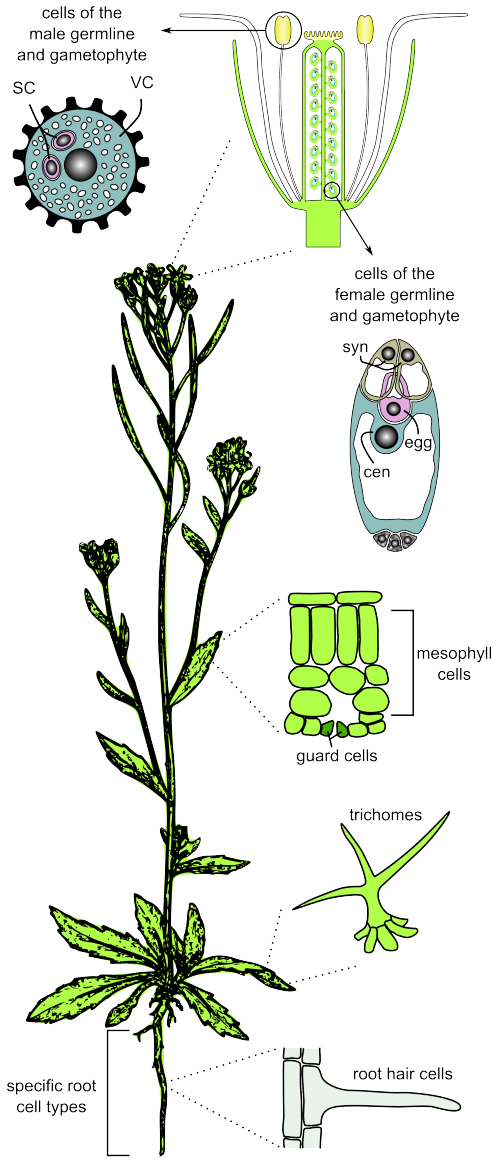


Figure 1. Cell and tissue types frequently used for cell type-specific systems biology and "omics" studies in plants. For the germlines, only the mature gametophytes are shown. Abbreviations: SC, sperm cell, VC, vegetative cell, syn, synergids, cen, central cell, egg, egg cell.

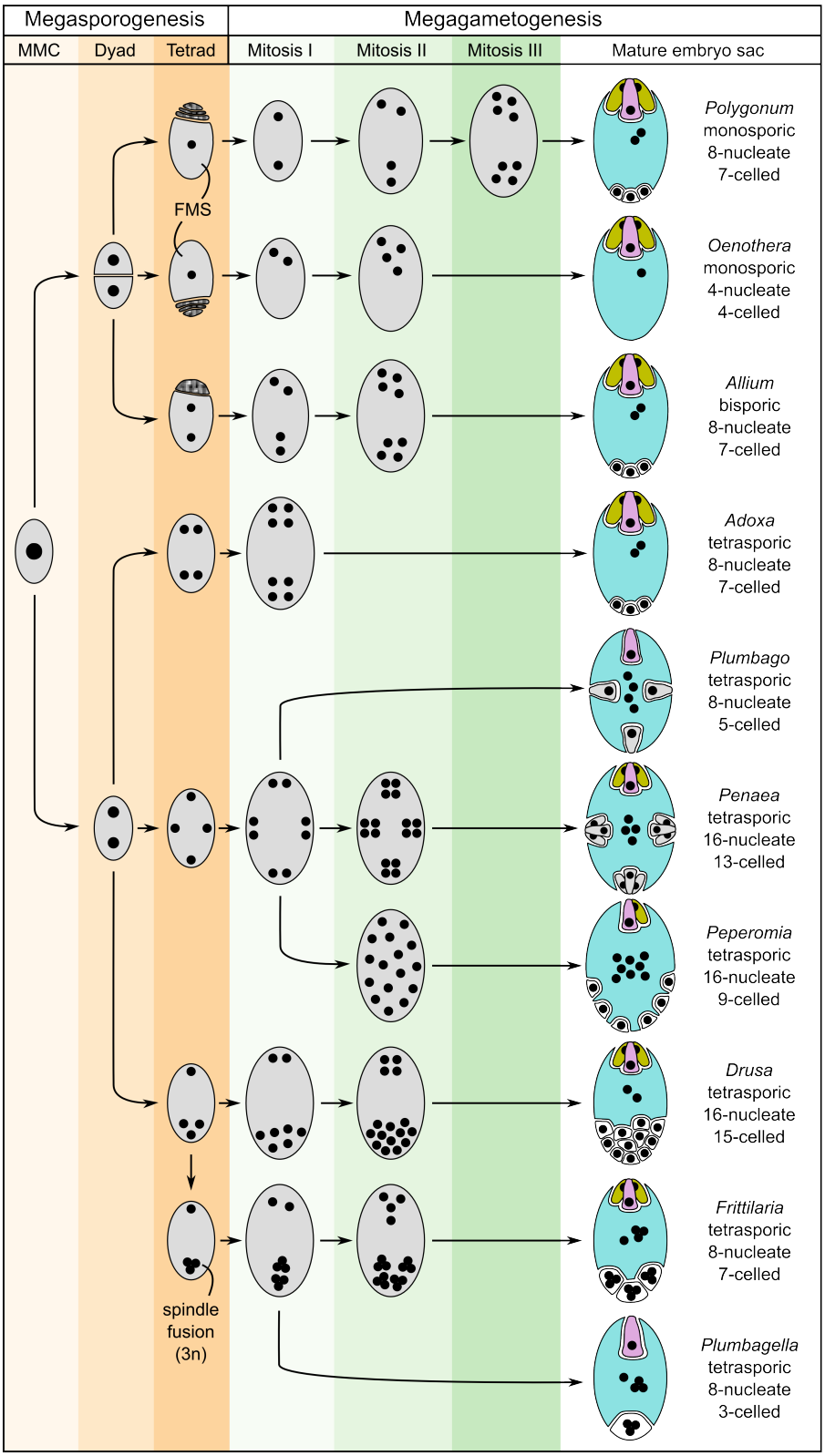

Megasporogenesis
Megagametogenesis
MMC
Dyad
Tetrad
Mitosis I
Mitosis II
Mitosis III
Mature embryo sac
FMS
spindle fusion (3n)
Polygonum monosporic 8-nucleate 7-celled
Oenothera monosporic 4-nucleate 4-celled
Allium bisporic 8-nucleate 7-celled
Adoxa tetrasporic 8-nucleate 7-celled
Plumbago tetrasporic 8-nucleate 5-celled
Penaea tetrasporic 16-nucleate 13-celled
Peperomia tetrasporic 16-nucleate 9-celled
Drusa tetrasporic 16-nucleate 15-celled
Frittilaria tetrasporic 8-nucleate 7-celled
Plumbagella tetrasporic 8-nucleate 3-celled

Figure 2. Schematic showing several basic types of female gametophyte development in angiosperms and the structural diversity of the mature embryo sacs (after [99]). The development of the female gametophyte can be devided into two steps: megasporogenesis (orange shading) and megagametogenesis (green shading). During megasporogenesis, a selected sporophytic cell, the megaspore mother cell (MMC), undergoes meiosis to give rise to spore formation. In most angiosperms, a tetrad of four megaspores is formed, of which three subsequently abort, leaving only one functional megaspore (FMS) to participate in megagametogenesis (e.g., *Polygonum*-type). However, a high diversity of the developmental processes of megasporogenesis and megagametogenesis has been observed in different genera, with variations for example including bispory and tetraspory. During megagametogenesis, the mature female gametophyte is formed through mitotic divisions, nuclear migration, and cellularization. For the mature embryo sac, the colors indicate the cell types: egg (pink), synergids (yellow), central cell (blue), and antipodal/lateral cells (white). Cells structurally similar to egg cells or synergids are drawn accordingly but colored in gray.

Tables

Table 1. Summary of transcriptome (top) and proteome (bottom) datasets generated for specific cell types during formation of the male reproductive lineage and gametogenesis. In brief, pollen formation starts with a microspore mother cell which undergoes meiosis to give rise to a tetrad of reduced spores. Each of these microspores undergoes pollen mitosis I to give rise to a generative and a vegetative cell. The subsequent mitotic division of the generative cell (pollen mitosis II) results in the formation of two sperm cells [93]. Abbreviations: PMC, pre-meiotic microspore mother cell, UNM, uninucleate microspore, GC, generative cell, SC, sperm cell, LC, liquid chromatography.

species	cell type	profiling method	literature
Oryza sativa ssp. *Japonica*	PMC	44K Agilent microarray	[139]
Zea mays	meiocyte	RNA-Seq (Illumina)	[96]
Arabidopsis thaliana	meiocyte	RNA-Seq (SOLiD)	[140]
Arabidopsis thaliana	meiocyte	RNA-Seq (Illumina)	[77]
Arabidopsis thaliana	meiocyte, UNM	CATMA microarray	[141]
Arabidopsis thaliana	UNM	Affymetrix ATH1 microarray	[142]
Oryza sativa ssp. *Japonica*	UNM	Affymetrix rice genome array	[143]
Lolium longiflorum	GC	cDNA microarray	[144]
Arabidopsis thaliana	SC	Affymetrix ATH1 microarray	[145]
Plumbago zeylanica	SC	cDNA spotted microarray	[146]
Oryza sativa ssp. *Japonica*	SC	RNA-Seq (Illumina)	[95]
Nicotiana tabacum	PMC, tetrad, UNM, polarized UNM	gel LC-MS	[70]
Solanum lycopersicum (ecotype Red Setter)	PMC, tetrad, UNM, polarized UNM	gel LC-Orbitrap-MS	[69]
Lilium davidii var. unicolor	SC, GC	MS/MS, MALDI-TOF/TOF	[97]
Oryza sativa ssp. *Nipponbare*	SC	LC-MS/MS	[68]

Table 2. Summary of transcriptome (top) and proteome (bottom) datasets generated for specific cell types during formation of the female reproductive lineage and gametogenesis. Abbreviations: MMC, megaspore mother cell, AIC, apomictic initial cell, AI, aposporous initial cell, egg, egg cell, syn, synergids, cen, central cell, LC, liquid chromatography.

species	cell type	profiling method	literature
Arabidopsis thaliana	MMC	ATH1 microarray	[56]
Arabidopsis thaliana	egg, cen, syn	ATH1 microarray	[55]
Arabidopsis thaliana	cen	RNA-Seq (SOLiD)	[57]
Arabidopsis thaliana	egg, syn	RNA-Seq (SOLiD)	[58]
Oryza sativa ssp. *Nipponbare*	egg, cen	44K Agilent microarray	[122]
Boechera gunnisoniana	AIC, egg, cen, syn	ATH1 microarray, RNA-Seq (SOLiD)	[58]
Hieracium praealtum	AI	RNA-Seq (Roche 454)	[53]
Oryza sativa ssp. *Nipponbare*	egg	LC-MS/MS	[68]

3 Polarized distribution of mRNA in the syncytial female gametophyte of *Arabidopsis thaliana* precedes cellularization and cell specification

The following manuscript is intended as a research article. Ueli Grossniklaus, Anja Schmidt, and I designed and conceived the study. I designed, generally performed, analyzed, and interpreted the results of the individual experiments. However, several experiments were carried out by the co-authors. Anja Hermann performed the seed set counting and the cytological characterization (images in, and data underlying figure 6 and 7). Daniela Guthörl carried out the RNA *in situ* hybridization experiments (Peter Schmid cloned some of the probes together with her). Afif Hedhly processed the final microscopy images (partly together with Anja Hermann) and wrote the paragraph "Microscopy figure processing". Stefan Grob contributed to the cloning of the construct for uniform overexpression. Ulrich C. Klostermeier performed the RNA-Seq library preparation, quality controls (data underlying Supplemental File S1), and the sequencing (supported by Philip Rosenstiel). I further analyzed all data, wrote the manuscript, and created/assembled all tables and figures. Anja Schmidt critically read the manuscript and provided valuable feedback.

Polarized distribution of mRNA in the syncytial female gametophyte of *Arabidopsis thaliana* precedes cellularization and cell specification

Marc W. Schmid[1], Anja Schmidt[1], Anja Herrmann[1], Afif Hedhly[1], Daniela Guthörl[1], Stefan Grob[1], Peter Schmid[1], Ulrich C. Klostermeier[2], Philip Rosenstiel[2], Ueli Grossniklaus[1,*]

1 Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zürich, Switzerland

2 Institute of Clinical Molecular Biology, Christian-Albrechts University, Kiel, Germany

*** E-mail: grossnik@botinst.uzh.ch**

Abstract

The female gametophyte (FG) is crucial for the reproduction of angiosperms. It harbors the two female gametes, the egg and central cell, which, following fertilization, give rise to the embryo and endosperm, respectively. Development of the *Polygonum*-type FG starts with a single haploid functional megaspore that undergoes three mitotic divisions in a syncytium. Nuclear migration and concomitant cellularization eventually leads to the formation of an eight-nucleate, seven-celled FG. It is a highly polarized structure containing four distinct cell types: the synergids, the egg cell, and the central cell at/towards the micropylar pole (the site of pollen tube attraction) and the antipodal cells at the opposite chalazal pole. It is still unclear how these cells are specified. We hypothesized that, in analogy to embryogenesis in *Drosophila*, specific subcellular localization of mRNA may be involved in controlling female gametogenesis in *Arabidopsis*. We therefore profiled the transcriptomes of the chalazal and micropylar halves of the syncytial FG using a combination of laser-assisted microdissection (LAM) and RNA-Seq. Comparative transcriptome analysis identified 615 genes displaying polarized localization of their transcripts within the syncytial FG. The data were validated using RNA *in situ* hybridization and reporter gene fusions. In addition, we screened for alterations in FG development and cell identity in a gain-of-function experiment, in which candidate genes were uniformly overexpressed during FG development. The data suggested that polarized localization of mRNAs might be involved in specifying cell-fate as well as controlling protein localization prior to translation, potentially to regulate the transition from the syncytial to the cellular stage. In summary, the data provide a basis for the identification of novel genes involved in FG development and may aid to study the role and mechanisms of subcellular mRNA localization in plants.

Introduction

Plants have a life cycle with an alternation between two heteromorphic generations: the sporophyte and the gametophyte. In the diploid sporophyte, distinct cells undergo meiosis and produce haploid spores. These give rise to multicellular, haploid gametophytes, which produce the gametes by mitotic division. Fusion of a male and a female gamete results in a zygote, from which the sporophyte is formed again. Because of its simple structure with few, yet highly distinct, cell types, the female gametophyte (embryo sac) of *Arabidopsis thaliana* is an attractive system to study basic principles of development [1]. Development of the female reproductive lineage starts with an archespore, which directly differentiates into the megaspore mother cell (MMC) [2]. Meiosis of the MMC results in a tetrad of haploid megaspores, of which only one, the functional megaspore (FMS), survives. Following only three rounds of mitosis in a syncytium, nuclear migration, and concomitant cellularization, the mature embryo sac consists of only four cell types: the antipodals, the synergids, the egg, and the central cell [3, 4] (Figure 1).

Up to now, several key processes and genes involved female gametogenesis have been identified in forward genetic screens or with reverse genetic approaches [1, 5]. During the syncytial phase, oriented mitosis and nuclear positioning are crucial. After the first mitosis, the two nuclei are placed on the opposite poles (chalazal and micropylar) and get separated by a central vacuole (Figure 1). In *Arabidopsis*, following the second mitosis, the nuclei are first positioned in transverse orientation to the chalazal-micropylar (long) axis and then start to migrate and become positioned along the long axis [3]. It seems that already at this four-nucleate stage, the fate of the nuclei is predetermined (Figure 1): The micropylar-most nucleus divides transverse to the long axis and gives rise to the two synergid nuclei. The second nucleus at the micropylar pole divides longitudinal to the long axis forming the egg cell nucleus and one of the two polar nuclei of the central cell. The nuclei at the chalazal pole divide in a similar pattern giving rise to three antipodal cell nuclei at the chalazal pole and the second polar nucleus of the central cell [6]. Interestingly, determination of cell fate depends on the position of the nuclei as for example indicated by the *Arabidopsis retinoblastoma-related 1 (rbr1)* mutant, which produces supernumerary nuclei differentiating according to their position within the FG [7]. This, and additional evidence from other studies (reviewed in [1]), points to the existence of two, at least partly, independent processes involved in cell specification: The positioning of the nuclei themselves, and the existence of information at a specific position within the syncytium. The nature of this latter positional information is still under debate. An appealing possibility for positional information are gradients of plant hormones, such as auxin or cytokinin, for both of which a role in establishing polarity during FG development has been proposed. Auxin has been implicated in the formation of a patterning gradient in the FG as shown by abolishing its potential gradient either by expressing the auxin biosynthethic protein YUCCA1 (YUC1) in the entire FG, or by dowregulating selected *AUXIN RESPONSE*

FACTOR (ARF) genes [8]. However, recent studies indicate that there is no auxin gradient in the FG; but auxin signaling is restricted to the surrounding sporophyte in a polar manner, which in turn may influence FG development in a non-cell-autonomous way [9, 10]. For cytokinin, an elevated signaling activity in the chalazal part of the ovule has been identified to be essential for FG development [11–13]. This process is likely regulated by both, the cytokinin-dependent *Arabidopsis* histidine kinases (AHKs) and the cytokinin-independent histidine kinase CKI1, which lacks a cytokinin-binding domain and activates cytokinin signaling autonomously [5]. In both cases, activation leads to a phosphotransfer to nuclear B-type response regulators (RR) via histidine phosphotransfer proteins (HPTs). The B-type RRs then induce transcription of a set of downstream genes as well as A-type RRs, which in turn repress cytokinin signaling (negative feedback) [14]. Due to genetic redundancy, the role of the individual RRs during female gametogenesis is unclear [13].

Aside plant hormones, RNA processing and metabolism seems to play an important role in the formation of the reproductive lineage and gametogenesis, as indicated by mutations in genes of the putative core spliceosomal components (*lachesis (lis)* [15], *clotho (clo)* [16], and *atropos (ato)* [17]), genes involved in biogenesis of ribosomal RNA (*slow walker1 (swa1)* [18]), or genes coding for RNA helicases (*slow walker3 (swa3)* [19], *mneme (mem)* [2], *suppressor of var 3 (suv3/eda15)* [20], and *magatama3 (maa3)* [21]). Interestingly, these mutations in genes involved in RNA metabolism also point out an alternative nature of the positional information: RNA might be specifically processed in and/or transported to certain sub-compartments of the syncytial FG. These transcripts may then serve as cell-fate determinants, allow an immediate translation following cellularization and cell specification, or control the subcellular localization of proteins they encode for prior to translation [22]. Little is known about subcellular mRNA localization in plants and only few examples are studied up to date [23]. An interesting example is the pollen of *Nicotiana tabacum*, in which some transcripts accumulated during early developmental stages, but were only translated at the time of pollen tube growth. Analysis of protein extracts further indicated that these late-translated transcripts were stored in messenger ribonucleoprotein particles (mRNPs) and sequestered from translation [24]. Subcellular mRNA localization is prevalent in animals and plays central roles in many cellular events [25]. Polarized mRNA localization was for example reported for over 70% of the genes expressed during early embryogenesis in *Drosophila*, with a peak around the transition from syncytial to cellular development potentially reflecting the high demand for localization events [22]. Considering that female gametogenesis in *Arabidopsis* as well involves syncytial development followed by cellularization and cell specification events, we hypothesized that subcellular localization of mRNAs may be involved in female gametogenesis in *Arabidopsis*.

To test the hypothesis that mRNAs are localized in a polarized manner within the syncytial FG, we studied and compared the transcriptional profiles of the chalazal and the micropylar poles of the FG at the late two- to early eight-nucleate stage (late FG3 to early FG5 [3], Figure 1). We therefore separately isolated and profiled the two cell halves of the syncytium with a combination of laser-assisted microdissection (LAM) and RNA-Seq [26]. Transcriptome analysis identified 615 transcripts which are enriched in one half of the FG. In a gain-of-function approach, we aimed to uniformly overexpress 34 of these genes in the entire FG throughout gametophyte development starting from the one-nucleate FMS (FG1 [3], Figure 1) stage on. We screened for or high percentages of ovule abortion and for alterations of cell fate using a plant line distinctly marking all four cell types of the mature FG [27]. Out of seven candidates identified, we characterized three in greater detail: *RKD2*, an egg-cell inducer [28], *MYB64*, a transcription factor required for FG development [29], and *ARR9*, an A-type response regulator of cytokinin signaling [14]. Taken together, the results suggested that subcellular localization of mRNA within the developing FG does occur and that it likely plays a role in FG development and cell type specification following cellularization. We believe that the transcriptional data generated in this study can serve as a basis to identify novel genes and molecular mechanism involved in FG development and has the potential to act as a starting point to study the processes involved in subcellular localization of mRNA in plants.

Results and Discussion

The transcriptome of the syncytial female gametophyte of *Arabidopsis*

To determine the transcriptional profile of the syncytial FG, we separatey isolated the two cell halves of the syncytial FG with a focus on the four-nucleate stage with laser-assisted microdissection (LAM, illustrated in S1 Figure, triplicates for each half). Given the asynchronous development of the embryo sacs within the ovules of the same flower (late FG3/FG4 and FG4/early FG5 stages often co-occur [3]), the small size of the cell halves, and the structural and optical limitations of dry thin sections required for LAM, it was difficult to differentiate between late two-nucleate, four-nucleate, and early eight-nucleate embryo sacs. To monitor potential cross-contamination with the surrounding sporophytic tissue, we isolated this as well separately (micropylar and chalazal surrounding sporophytic tissue, one sample for each). The gametophytic samples were always harvested before the sporophytic tissues and only well visible structures were selected, arguing for a high purity of the gametophyte samples. In agreement, the high similarity within the gametophytic and within the sporophytic samples (Figure 2A, S2 Figure) indicated that the cross-contamination of gametophytic samples with the surrounding sporophytic tissue was likely marginal.

Following the isolation of RNA, libraries were prepared as described previously [26] and sequenced on one eights of a SOLiD slide resulting in 36 to 50 million reads per sample (Table 1). Reads were corrected for potential sequencing errors with the SOLiD Accuracy Enhancement Tool (version 2.2, available via www.biostars.org/p/52250) and aligned to the reference genome of *Arabidopsis thaliana* with Subread [30] resulting in around 5 to 18 million weighted alignments (in the following termed "hits", see [26, 31] for details, Table 1). In addition to the data generated in this study, we processed data from several different tissue and cell types (see Material and Methods for details, S1 Table, S2 Table). To get a genome-wide overview of the transcriptional profiles, we classified the hits based on their genetic context (Table 1, S3 Table) and compared all samples to each other (S3 Figure). All samples generated in this study were highly similar to each other and to the majority of publicly available data with around 90% of all hits falling into known exons or covering known splice junctions. Together with the high reproducibility of on average above 0.85 (Figure 2A), this indicated a generally high quality of the transcriptional profiles.

To get an estimate of the transcriptome size, we defined genes having at least 25 hits in two out of three replicates as being expressed in the respective half of the FG. Applying this (arbitrary) threshold, we found 14'367 and 14'179 genes to be expressed in the micropylar and the chalazal half of the syncytial FG, respectively (15'660 genes in total). This was slighly less compared to more complex tissue types (e.g. 16'769 in roots and 16'333 in seedlings [32], defined as at least 25 hits in three out of five replicates). This was likely due to the higher heterogenity in these tissues compared to individual halves of only one differentiating cell. However, it should be noted that the number of genes identified to be expressed in a certain sample can vary as well due to the differences in the number of replicates and the library preparations.

Functional categorization of the 15'660 genes expressed within the syncytial FG and comparison of Gene Ontology (GO) annotation to the entire set of transcripts of *Arabidopsis* results in 180 significantly enriched GO terms (belonging to the domain "biological process"). Most of them were per default overrepresented in any transcriptome irrespective of the tissue or cell type under investigation. To filter for the specific terms, we compared the transcriptomes of all available tissue and cell types (S1 Table, except cell types of the mature FG) to the entire set of transcripts of *Arabidopsis* and removed all GO terms which were found to be significant in any of these comparisons. Out of the initial 180 GO terms, only 3 passed this filter: "tetrapyrrole metabolic process" (GO:0033013), "regulation of cell division" (GO:0051302), and "double fertilization forming a zygote and endosperm" (GO:0009567) (S4 Table). The first term (GO:0033013), to speculate, might suggest increased retrograde signaling from the plastids to the nucleus or ABA signaling [33]. The second term (GO:0051302) was likely indicative of the highly regulated mitotic pattern during FG development and suggested that some genes impor-

tant for cellularization are already expressed during the syncytial phase. The last term (GO:0009567) comprised many genes which either play a role during female gametogenesis (e.g. *RETINOBLASTOMA RELATED (RBR1)* [7], *VERDANDI (VDD)* [34]) or are important for specific functions of the mature FG (e.g. *LORELEI (LOR)* [35]). Similarly, when comparing a list of genes known to be involved FG development and function to the list of expressed genes, most of them were expressed (35/41, S5 Table). Interestingly, some of them were only described to be characteristic for a specific cell type of the mature FG (e.g. *DIANA (DIA)* [36, 37], and *MYB98* [38]), supporting the notion that, at least for some genes, transcripts required in the mature FG are produced already earlier during female gametogenesis.

To identify individual genes specifically involved in female gametogenesis during the syncytial stage, we compared the samples of the syncytial FG to all other samples processed in this study (S1 Table, including the samples of the mature FG). Out of the 15'660 genes expressed within the syncytial FG, 101 were highly enriched (FDR < 0.05 and logFC > 5). Functional categorization revealed that almost half (46) of the genes encoded for extracellular proteins, of which most were small cystein-rich proteins and cysteine-rich receptor-like kinases (CRK/DUF26) (Figure 2B, S6 Table). However, their function is largely unknown. They may play a role in apoplastic cell-to-cell communication of the FG with the surrounding sporophyte given that symplastic connections between the FG and the surrounding sporophytic cells are likely absent [39].

Considering that previous transcriptome studies on the female germline of *A. thaliana* [2, 40] were done with the ATH1 microarray, on which only around two third of the transcriptome are represented with a probeset [26], we compared the RNA-Seq samples of the FG (micropylar and chalazal half of the syncytial FG, central cell [26], synergids, and egg cell [41]) to all other data processed in this study (S1 Table). We could identify 363 genes highly enriched (FDR < 0.05 and logFC > 5) in the FG. Interestingly, this list was dominated by genes specific to the cell types of the mature FG, which formed a well defined cluster (S4 Figure). Surprisingly, both halves of the syncytial FG were distinct from the cell types of the mature FG. This indicated that only a subset of genes important for distinct cell types of the mature FG were already expressed within the syncytial FG (above). Many others seem to be *de novo* expressed upon cellularization or later on after cell determination.

Functional categorization of the genes specifically expressed within the FG using GO-term enrichment identified small nucleolar RNAs (snoRNAs) as highly enriched in the FG (S7 Table). Interestingly, illustration of all 71 snoRNA genes of *A. thaliana* (TAIR10 annotation) revealed an almost exclusive enrichment of snoRNAs in the mature FG and the male meiocytes [42] (S5 Figure). This observation was unlikely due to a technical artifact linked to library preparation as this would have affected the samples of the syncytial FG

and the surrounding sporophyte as well. It should be noted, however, that this may be an effect of the limited number of tissue types with RNA-Seq data available (none of the snoRNAs is represented on the ATH1 microarray). Nonetheless, this strong enrichment points out the importance of RNA metabolism and, specifically rRNA biogenesis for FG development and function.

Polarized localization of mRNA within the syncytial female gametophyte

To test whether some transcripts are more abundant in one half of the syncytial FG, we compared the transcriptional profiles of the chalazal and the micropylar poles with each other. It should be noted that this tests for genes whose transcripts are localized in a polarized manner. This may be achieved by (i) differential expression, (ii) active mRNA-transport, (iii) anchoring of passively diffusing RNAs, or (iv) localized degradation of certain RNAs. In the following sections, we will therefore use "polarized genes" as generic term for all four possibilities. Using edgeR [43] with "trended" dispersion estimates, we could identify 615 polarized genes, out of which 554 and 61 were enriched in the micropylar and chalazal half, respectively (Figure 2C,D, S8 Table). It is possible that the difference in the number of enriched genes reflects to a certain extent the slightly higher quality of the micropylar samples. However, it is more likely due to the difference in the complexity of their lineage. The micropylar half gives rise to three very distinct cell types, whereas the chalazal half only contributes to two, out of which one seems not to be of major importance for the function of the mature FG.

The polarized genes were mostly protein-coding (595). Only few genes were classified as transposable element gene (10), pseudogene (5), or genes coding for "other RNA" (3), a U6 small nuclear RNA (1), or plastidial ribosomal RNA (1, S8 Table). Functional categorization of the polarized genes using GO-Slim terms revealed a high functional diversity and/or limited annotation (Figure 2E). Most of the genes were classified into broad terms (within "biological process": "other cellular processes" and "other metabolic processes"; within "molecular function": "other binding" and "unknown molecular functions"). In respect to the GO-domain "cellular component", the genes were diverse as well, however, with the largest groups falling into "nucleus", "extracellular", or "other cytoplasmic components" (Figure 2E). In-depth GO-term enrichment analysis using topGO [44] identified several interesting terms enriched in the set of polarized genes compared to the set of genes tested for polarized expression (S9 Table). The most significant term "pollen tube development" (GO:0048868) seemed unusual, however, it includes several genes known to be important for FG development and function (e.g., *MAA3* [21], *MYB98* [38], *LOR* [35], and *UNE9* [20]). Enrichment of "histone methylation" (GO:0016571) potentially indicated transcriptional regulation, or the establishment of specific epigenetic states. Other terms, such as "syncytium formation" (GO:0006949), "actin filament-based

movement" (GO:0030048), "cell wall modification" (GO:0042545), and "cell-cell signaling" (GO:0007267), may reflect the stringent regulation of cell division and organization in the syncytial FG. Subcellular localization of these mRNAs may act to control the localization of the protein prior to translation. Potential advantages of transporting mRNA instead of proteins are cost effectiveness and limitation of protein activity to the desired compartment within the cell: a single mRNA molecule can be used to produce many proteins, thereby avoiding the need to transport each protein individually and to prevent it from acting during its transport [22].

Notably, neither auxin nor cytokinin signaling were identified as being enriched in the list of polarized genes using GO-terms. However, it is possible that the responses would not be visible as a strong change of several individual genes, but as a concerted change of expression of a larger set of genes acting together. We therefore tested whether the entire set of auxin/cytokinin responsive genes exhibited polarization in the syncytial FG using the Gene Set Enrichment Analysis (GSEA) approach described in [45, 46]. Genes tested for auxin comprised the 29 members of the *Aux/IAA* and the 23 members of the *ARF* gene families in *Arabidopsis*. Cytokinin responsive genes (226) were taken from the "golden list" described in [47]. Genes were filtered for expression within the developing FG (defined as having at least 25 hits within two out of three replicates of any half of the syncytial FG). Out of the 52/226 genes in total, only 20/73 genes were considered to be expressed and passed the filter. GSEA on these genes revealed a highly significant enrichment of cytokinin-responsive genes in the chalazal half of the syncytial FG (P-value < 0.0001, S6 Figure). This is in line with reports identifying cytokinin signaling activity in the chalaza of the ovule [11–13]. Auxin-responsive genes displayed a random distribution (P-value $= 0.374$). Nonetheless, transcripts of two auxin response factors (ARF9 and ARF18) were significantly enriched in the micropylar half of the syncytial FG. Thus, it may be that auxin signaling in the FG is highly specific, involving only few genes. However, the importance of these two ARFs remained unclear, given that none of them was specifically targeted by the artificial microRNA in [8].

To validate the data, we tested the expression of individual candidate genes with two independent approaches: monitoring of mRNA localization with RNA *in situ* hybridization, and visualization of promoter activity using reporter constructs (nuclear localized red fluorescent protein as transcriptional reporter). We could confirm the polarized localization of mRNA for two genes tested with RNA *in situ* hybridization (Figure 3, Table 2). Reporter constructs for putative promoters of most genes enriched in the micropylar half were active in the synergids, the egg cell, or the egg apparatus and the central cell, partly confirming the data (Figure 3, Table 2). However, we rarely detected a fluorescent signal at stages earlier than the eight-nucleate (FG5) stage. It was unclear, whether this may have been due to a translational delay or missing control elements in the putative promoter sequences (or the absence of the gene sequences themselves).

Uniform overexpression of candidate genes partially leads to cell type mis-specification in the mature FG

To functionally validate the data and to identify potential candidates involved in FG development and specification, we chose a gain-of-function approach in which we aimed to distribute polarized mRNAs equally within the entire FG. We therefore expressed candidate genes under the control elements of *AT4G05440* (promoter and UTR with terminator), which were shown to drive expression of reporter genes in the entire FG starting from FG1 on (*ANIKEVORKIAN (AKV)* reporter in [48], and *AtD123* reporter in [49]). We removed the 5' and 3' untranslated regions (UTRs) of the candidate gene sequences (genomic) as these might have acted as anchor points for RNA-binding proteins tethering the mRNAs to their regular location [23]. Candidates were screened for alterations of cell fate and high rate of ovule abortion (qualitative observation) using a plant line allowing to distingiush all four cell types of the mature FG (termed quadruple marker, Figure 4) [27]. Out of seven candidates identified in the screen (Table 3), we characterized three in greater detail: *RKD2*, an egg cell inducer [28], *MYB64*, a transcription factor required for FG development [29], and *ARR9*, an A-type response regulator of cytokinin signaling [14].

RKD2 is sufficient to induce egg cell-like fate in the female gametophyte

RKD2 is a transcription factor of the RKD transcription factor family with the potential to induce an egg cell-like transcriptome in somatic tissues [28]. However, due to genetic redundancy and the choice of the promoter for overexpression in the latter study, its role during female gametogenesis remained speculative [5]. In this study *RKD2* was identified as a polarized gene displaying a significant enrichment in the micropylar half of the syncytial FG (S8 Table). We therefore tested whether uniform expression during the entire female gametogenesis would result in an alteration of cell fate (i.e. the induction of ectopic egg cell-fate). Out of eight independent T1 plants carrying the construct for uniform overexpression of *RKD2* (construct ID "ox13"), four showed consistent conversion of the synergids and the central cell into cells expressing the egg cell marker (Table 3, Figure 4D, Figure 5). In most cases, cell shape and position were strongly distorted as well (Figure 5A,B). Among the approximately 150 ovules displaying an alteration in marker expression during the screen, most displayed expression of the marker in almost all synergids and central cells (only in four cases, expression of the synergid marker could still be observed in one of the synergids) (Figure 5C). In contrast, expression of the antipodal marker was frequent (Figure 5A). Conversion rates were close to the expected 50% for a case with a single/linked transgene insertion in T1, with on average 45% of all FGs (N = 448) displaying multiple cells expressing the egg cell marker (two independent lines, 42.6% and 48.3%, Figure 5C). To test whether the expression of the egg cell marker at the position of the synergids was accompanied by a loss of synergid function, we assessed

if the ovules were still capable of attracting pollen tubes. Pollen tube reception was strongly reduced in lines carrying the ox13 construct compared to the quadruple marker background and not observed in around 56% of the ovules (N = 1'585, P-value < 0.001, Figure 5D). This suggested that *RKD2* was not only sufficient to induce egg cell marker expression, but to change the synergids sufficiently to abolish their designated function. Interestingly, even though *RKD2* should have been expressed from the FG1 stage on, conversion occurred only late during female gametogenesis. It may imply that either a certain number of mitotic divisions, or cellularization were required for egg cell-fate. However, the possibility that a remaining control element in an intron of the construct inhibited early activity can not be excluded.

The formation of multiple egg cells has recently been reported in plants carrying a mutation in the gene *ALTERED MERISTEM PROGRAM 1 (AMP1)* [50]. Interestingly, only homozygous, but not heterozygous, *amp1* plants showed the formation of multiple egg cells at the cost of the synergids. This indicated that an AMP1-dependent signal from the surrounding sporophyte could promote or maintain synergid cell fate. Intriguingly, specific expression of *AMP1* in either synergids, the central cell, or even the egg cell of homozygous mutant plants could restore the phenotype [50]. The latter implies that AMP1 (or a derived signal) was not sufficient to induce synergid fate on its own (otherwise the egg cell would have been affected upon expression of *AMP1*). The strong and highly penetrant conversion of synergids upon *RKD2* overexpression furthermore suggests that *AMP1*-mediated maintenance of synergid cell-fate can be overridden by *RKD2*. This indicates that *RKD2* may be an upstream regulator of egg cell-fate. *RKD2* might therefore be an example of a cell fate determinant in the FG whose activity is regulated by specific localization of its mRNA.

Various effects of *MYB64* on female gametogenesis

MYB64 is a member of the MYB transcription factor family. Together with *MYB119*, it has recently been shown to be important for the transition of the syncytial FG5 stage to the cellularized FG6 stage (FG5 transition, [29]). Both, *MYB64* and *MYB119*, were identified as polarized genes with a significant enrichment in the micropylar half of the syncytial FG (S8 Table). We tested both for alteration of cell fate upon uniform overexpression during female gametogenesis (Table 3). However, given the similarity in the screen (Table 3) and their genetic redundancy [29], we only characterized *MYB64* in greater detail. Out of 13 independent T1 plants carrying the construct for uniform overexpression of *MYB64* (construct ID "ox10"), six showed irregularities during FG development (Table 3). Either synergids or the egg cell appeared to be sometimes not specified (Figure 4E) and the rate of ovule abortion seemed to be increased. However, conversion rates appeared to be low and quantification with confocal microscopy was not feasible. We therefore first tested whether the seed set was affected upon uniform overexpres-

sion of *MYB64*. The percentage of normally developing seeds was indeed significantly reduced by around 30% (N = 5'106) compared to the quadruple marker background (two independent lines, 27.6% and 34.1% reduction, Figure 6H). There was an increase in both, unfertilized/pre-fertilization aborted ovules as well as post-fertilization aborted seeds. This suggested a variable effect of uniform overexpression of *MYB64* with some FGs still developing normal enough to attract pollen tubes and to be fertilized. Accordingly, cytological characterization of the two lines with a reduced seed set revealed a large phenotypic variation during FG development. We summarized the phenotypes into three classes (Figure 6B-F): (I) Seven individual cells were recognizable and the polar nuclei were clearly distinct from the other nuclei. Individual cells may have been slightly misplaced and polar nuclei remained unfused (Figure 6B). (II) Seven or less individual cells were recognizable, however, nuclei were frequently severely misplaced and their identity could not be distinguished. Cells may have been extremely small (Figure 6C-E). (III) FG was completely degraded/missing (Figure 6F). In both lines we could detect a significant decrease in normal mature FGs of around 25% (N = 1'160) and a significant increase of phenotypes belonging to class II and III. Phenotype I, which was mainly defined by unfused polar nuclei was slightly increased compared to the quadruple marker background, but differences were not significant (Figure 6G).

The presence of both, unfertilized/pre-fertilization aborted ovules as well as post-fertilization aborted seeds, and the highly variable FG phenotypes (especially class II comprised very diverse phenotypes, Figure 6C-E) may have reflected the various pathways *MYB64* feeds into the development of the FG and the transition from syncytial to cellular growth. Analysis of *myb64 myb119* double mutant plants revealed an involvement in cellularization and polarization of the FG, as well as the regulation of Fertilization Independent Seed (FIS) Polycomb Responsive Complex 2 (*PRC2*) required for seed development [29]. For some of the ovules belonging to the class II phenotypes, it might have been premature cellularization mediated by precocious expression of *MYB64* which lead to a developmental arrest (Figure 6E). In the case of post-fertilization aborted ovules, one could speculate whether misregulation of *FIS2* resulted in a defect during seed development. Finally, the importance of *MYB64* and *MYB119* for the micropylar cell identity, illustrated by the absence or severe reduction of the micropylar cell markers and an expanded expression of the chalazal cell marker in the *myb64 myb119* double mutant [29], coincides with the enrichment of their mRNAs at the micropylar pole (and the almost complete absence at the chalazal pole with only 11 and 12 hits in one out of three chalazal samples S8 Table). In contrast to this support of the importance of polar mRNA localization for cell type specification at the micropylar pole, we could not detect any induction of expression of the synergid, egg, or central cell markers at the chalazal pole upon uniform overexpression of *MYB64* or *MYB119*.

ARR9 supports a role for FG-autonomous cytokinin signaling during female gametophyte development

ARR9 is an A-type response regulator (RR) of cytokinin signaling [14]. It was identified as polarized gene with a significant enrichment in the chalazal half of the syncytial FG (S8 Table), where cytokinin signaling is likely important for FG development [11–13]. Out of twelve independent T1 plants carrying the construct for uniform overexpression of *ARR9* (construct ID "ox16"), six exhibited irregularities during FG development (Table 3, Figure 4F). The antipodal marker was frequently expressed at the central cell position and the cell appeared to be shifted towards the chalazal pole of the FG (Figure 4F). Additionally, ovule abortion appeared to be increased. As for ox10, conversion rates seemed too low to allow quantification with confocal microscopy. We therefore first tested for the reduction of seed set upon uniform overexpression of *ARR9*. Seed set was indeed significantly reduced on average by 19% (N = 5'368) compared to the quadruple marker background (two independent lines, 7.1% and 31.0% reduction, Figure 6E). The reduced seed set was mainly due to an increase in unfertilized/pre-fertilization aborted ovules suggesting a defect during FG development. By cytological characterization of the two lines with a reduced seed set, we could identify three different phenotypic classes (Figure 7A-C): (I) FG appeared normal but polar nuclei remained unfused and antipodal cells were sometimes shifted towards the central cell/polar nuclei (Figure 7A). (II) Certain cell/nuclei types, i.e. mainly the central cell nucleus and the antipodal nuclei, were not distinguishable anymore and were frequently equally-sized (Figure 7B). (III) FG was completely degraded/missing (Figure 7C). In both lines we could detect a significant decrease in normal mature FGs of around 22.3% (N = 618) and 59.7% (N = 599) and an increase of all three phenotypic classes. However, these increases were only significant in one of the two lines (Figure 7D).

Interestingly, the class I phenotype (unfused PN and sometimes misplaced AP) seemed to resemble the most frequent phenotype described for *cki1* mutant ovules [12]. Given that *ARR9* is a negative regulator of cytokinin signaling and *CKI1* encodes for a cytokinin-independent histidine kinase activating cytokinin signaling autonomously (i.e. without binding a cytokinin molecule), this suggested that the phenotype caused by uniform overexpression of *ARR9* was likely caused by distortion of cytokinin signaling. Together with a recent report showing specific expression of *CKI1* within the FG (with an enrichment towards the chalazal pole in later stages) [29] this observation thus supports the possibility of a cell-autonomous role of cytokinin signaling during FG development. Overexpression of ARR9 in the whole gametophyte migth thus reinforce the downregulation of cytokinin signaling in the chalazal half of the gametophyte. The additional expression of *ARR9* in the micropylar half of the ovule was likely irrelevant considering that cytokinin signaling seems to mainly act in the chalazal part of the ovule and the FG (see transcriptome above, [12, 13]).

Conclusion

We hypothesized that specific subcellular localization of mRNA may be involved in controlling female gametogenesis in *Arabidopsis thaliana*. We therefore profiled and compared the transcriptome of the micropylar and the chalazal halves of the syncytial FG from the late FG3 to the early FG5 stage [3]. We identified 615 genes for which the transcripts showed an enrichment in one of the two halves of the syncytial FG. Most of them were protein-coding. Among them were genes important for FG development and function and genes acting in syncytium formation, actin filament-based movement, and cell wall modification. We validated the expression patterns of several genes using RNA *in situ* hybridization and transcriptional reporters. To test for the functional implication of the polar RNA localization and to identify genes involved in FG development and cell-specification, we aimed to uniformly overexpress 34 candidate genes in the entire FG throughout its development. Screening for cell fate alterations and high rates of ovule abortion, we identified seven potential candidates, three of which we characterized in greater detail. One of them was *RKD2*, a gene known to have the potential to induce egg cell-like fate in vegetative tissues. However, its role during FG development remained unclear [5]. We observed that uniform overexpression of *RKD2* was sufficient to induce egg cell marker expression in the cells usually having synergid and central cell identity and to abolish synergid function. However, even though induction of egg cell-like fate was strong and highly penetrant, we did not observe a single case where egg cells were induced early during FG development. It appeared that either several rounds of mitosis and/or cellularization were prerequisites of egg cell-fate. However, we could not exclude the possibility that the late activity was due to a remaining control element in one of the introns of the transgene acting as translational suppressor.

In conclusion, our study indicates that polarized localization of mRNA during syncytial development of the FG of *Arabidopsis thaliana* precedes cell fate decisions upon cellularization. However, whether these mRNAs directly act as cell-fate determinants, allow a rapid translational burst after cellularization, or if subcellular localization serves as a mechanism to simply control protein localization prior to translation remains an open question. Whereas *RKD2* might be an example for a cell-fate determinant, the genes belonging to the GO-terms "actin filament-based movement" and "cell wall modification" are more likely to illustrate cases where specific mRNA localization is used to target proteins to the corresponding subcellular localization. However, clarifying the functional mechanism linking subcellular mRNA localization to protein localization and functional importance for cell type specification in the FG will further require a series of challenging experiments. Only recently, a tool for efficient *in vivo* monitoring of the exact subcellular localization of RNA in plants became available and it is not yet clear whether this method could be used to track the transcripts within the FG [23]. Importantly, to unambiguously describe the contribution of the mRNA to the localization of its protein,

it would in principle be necessary to monitor the protein independently of its mRNA for example by injecting a protein-reporter fusion. Aside the function of the polarized mRNAs themselves, the process controlling subcellular localization of RNA in plants is still largely unknown. We anticipate that the transcriptomes generated in this study may not only provide a basis to identify novel genes involved in FG development but also to study processes involved in subcellular localization of mRNAs in plants.

Materials and Methods

Plant material

Arabidopsis thaliana was grown under standard conditions as described previously [26]. Unless further specified, plants were of accession Landsberg *erecta*. Quadruple (single construct with embryo marker for selection: p*DD31:AcGFP1*, p*DD65:AmCyan*, p*DD45:DsRed-Express*, p*DD1:ZsYellow1*, and p*GmKTI3:AcGFP1* [27]) and p*AKV:H2B-YFP* [48] marker lines were kindly provided by S.J. Lawit and W.C. Yang, respectively.

Laser-assisted microdissection and RNA-sequencing

Inflorescences were fixed in 3:1 ethanol:acetic acid (v/v), vacuum infiltrated twice for 15 min on ice, and stored on ice over night before replacing the fixative with 70% ethanol. Flowers for the isolation of tissue from ovules at the four-nucleate embryo sac stage were selected based on morphological criteria and the developmental stage of the embryo sacs of the ovules in the tip of the ovary: From three flowers around the flower developmental stage 12 [51] of an inflorescence the tips of the ovaries were removed and transferred to clearing solution (chloral hydrate, water, and glycerol 8:2:1, w/w/w). This procedure was carried out under a binocular on a cold plate using syringes with BD microlance^{TM} 3 needles (BD, Franklin Lakes, USA). The rest of the flowers was stored individually in 70% ethanol at 4 °C. The cleared tips of the ovaries were examined under a differential contrast microscope (Leica DMR microscope, Leica Microsystems GmbH, Wetzlar, Germany). In case the tip contained ovules with an embryo sac at the two (with vacuole) or four-nucleate stage, the corresponding flower was transferred to a Leica ASP200 Tissue Processor (embedding machine). Embedding, sample preparation, and quality controls for LAM were done as described [26]. Embryo sac halves and the surrounding sporophytic tissue were sequentially isolated from 7 μm thin sections (illustrated in S1 Figure). Around 500 (micropylar) and 600 (chalazal) cuts of the embryo sac halves and 450 cuts of the sporophytic controls were pooled during the RNA extraction as described [26]. Library preparation, quality controls (S1 File), and SOLiD sequencing was done as described previously [26], except that the second and third replicates of the embryo sac halves were paired-end sequenced on SOLiD V4 instead of single-end on SOLiD V3. Raw data were deposited on SRA (SRP045521).

RNA *in situ* hybridization and promoter-reporter constructs

Genes for data confirmation by *in situ* hybridization and promoter-reporter gene constructs were selected based on (i) significant enrichment in one of the embryo sac halves, and (ii) expression values representing different expression levels. RNA *in situ* hybridizations experiments were performed on 8 μm thin sections as described previously [2] (see S2 File for a list of primers used for probe cloning).The putative promoter sequences of the selected candidate genes (i.e. the intergenic sequences 5'-upstream of a gene with a maximum length of 2 kb) were isolated from genomic DNA using primers containing adapters for ligation-independent cloning (see S2 File). The resulting PCR products were cloned into the vector pPVL11 containing a nuclear red fluorescent protein (SV40-tdTomato) as described in [52]. Plants were transformed following a simplified *Agrobacterium*-mediated plant transformation procedure [53]. Plants carrying the reporter constructs were crossed to the p*AKV:H2B-YFP* embryo sac marker line [48] (except constructs for *AT2G28400* and *AT1G70540*).

Uniform overexpression of genes with polarized localization of transcripts

Genes for uniform overexpression within the whole embryo sac were selected based on (i) significant enrichment in one of the embryo sac halves, (ii) preferentially specific expression in the female gametophyte compared to other tissues (whole plant and seedlings, unopened flowers, early globular embryos, male meiocytes, and 2-4 cell and globular stage embryos from [54], [55], [56], [42], and [57]), and (iii) stable RNA-Seq coverage pattern (visual inspection in GenomeView [58]). Truncated genomic gene sequences (i.e. without UTRs) were isolated from genomic DNA using primers with *attB1* and *attB2* sites (see S2 File), and cloned into the vector pDONR207 following the manufacturer instructions (Gateway Cloning, Invitrogen, Carlsbad, USA). Entry clones were recombined into the destination vector pMWS14, which drives embryo sac specific expression of the gateway-casette under the control of the promoter and the terminator regions of *AT4G05440* (same as *AKV* in [48], and *AtD123* in [49]). To generate the pMWS14 vecotr, AKV promoter and terminator regions were cloned into pMDC99 [59] using the *Asc*I and *Pac*I restriction enzymes, respectively (see S2 File for primer sequences). Quadruple [27] marker lines were transformed following a simplified *Agrobacterium*-mediated plant transformation procedure [53]. Gametophytes were screened for alterations in cell-fate (quadruple marker) two days after emasculation using confocal microscopy (for details see below; 30-50 randomly chosen gametophytes per plant, 8 to 10 T1 plants per construct). Carpel walls were removed and ovules were mounted on microscopy slides in MS-glycine solution (0.11 g MS salts and 3.75 g glycine in 50 ml water). Images were acquired using a Confocal Laser Scanning Microscope (Leica SP5-R, Leica Microsystems GmbH, Wetzlar, Germany) using a 63x glycerol immersion lens. If multiple fluorescent proteins were monitored, im-

ages were aquired sequentially (except in a first batch of the screen where all fluorescent proteins were imaged in parallel). Excitations/emissions for AcGFP1, AmCyan, DsRed-Express, ZsYellow1, tdTomato, and YFP were set to 488/500–525 nm, 458/465–500 nm, 561/575–625 nm, 514/520–540 nm, 561/584–636 nm, and 514/520–565 nm (AcGFP1, AmCyan, DsRed-Express and ZsYellow1 in the first batch: 488/493–509 nm, 458/460–485 nm, 561/527–560 nm, and 514/578–639 nm).

Cytological characterization

For cytological characterization, whole inflorescences and pistils two days after emasculation were fixed and examined under a differential contrast microscope as described for pre-screening of LAM material. Only plants displaying normal vegetative growth were used (the constructs ox10 (*MYB64*) and ox13 (*RKD2*) sometimes induced strong vegetative phenotypes).

Pollen tube reception assays

For characterization of pollen tube reception, plants were pollinated with pollen from the quadruple marker line two days after emasculation and prepared one day later for aniline blue staining of callose in pollen tubes as described [60]. Stained siliques were imaged with an epifluorescence microscope (Leica DM6000B, Leica Microsystems GmbH, Wetzlar, Germany). Only plants displaying normal vegetative growth were used (the constructs ox13 (*RKD2*) sometimes induced strong vegetative phenotypes).

Microscopy figure processing

Microscopy images, when needed, were corrected with noise reduction (4-6 intensity level) and/or Gaussian filter (radius 0.8-1.2 pixels) using Adobe Photoshop, Lightroom 4 and Adobe Photoshop CS5.1. DIC and fluorescence images (reporter constructs) were overlaid using Adobe Photoshop CS5.1 (subtract and clarify layer blending modes). All image manipulations were applied to the whole figure. Fluorescence and DIC image stacks were rendered using maximum intensity projection, and 2D still images were saved using the snapshot option (Imaris V7.0, Bitplane).

Data processing

RNA-Seq raw data processing

Short reads generated in this study were deposited at NCBI Sequence Read Archive (SRA, www.ncbi.nlm.nih.gov/sra) and are accessible through the accession number SRP045521. The transcriptomes generated in this study were compared to publicly available RNA-Seq transcriptome data from various tissues and cell types of *Arabidopsis thaliana*. The data

comprised roots, seedlings, and floral buds [32], male meiocytes [42], 2-4 cell and globular stage embryos [57], early globular embryos [56], endosperm and embryos at the torpedo stage [61], shoot apical meristems [62], pollen [63], central cells [26], egg and synergids [41], root non-hair and quiescent center cells [64], inflorescences, leaves, siliques [65], and flowers, leaves, and seedlings [66–68] (see S1 Table for accession numbers). Only data from untreated wild-type plants were used in the analysis. SOLiD reads were processed with the SOLiD Accuracy Enhancement Tool (version 2.2 with a reflength of 13'000'000 and the option -qvupdate). Reads were filtered for ribosomal RNA sequences with filterReads (github.com/MWSchmid/filterReads, only Illumina reads) and then aligned with Subread (i.e. subjunc, version 1.4.5; [30]) allowing up to ten alignments per read and with the option --allJunctions enabled. Count tables were created with Rcount [31] as described [26] but using the read length as allocation distance for calculating the weights of the reads with multiple alignments. Count tables were equalized with edgeR (version 3.8.5, [43]).

Differential expression and specific expression

Genes differentially expressed between the chalazal and micropylar halves of the four-nucleate embryo sac or between the four-nucleate embryo sac/female gametophyte samples and other tissue/cell types (see S1 Table) were identified with edgeR (version 3.8.5, [43]) using trended dispersion estimates and Benjamini-Hochberg multiple testing corrections. Genes with an adjusted *P*-value (FDR) below 0.05 and a minimal logFC of 2 (within gametophyte comparison) and 5 (gametophyte compared to other tissues) were considered to be differentially expressed. Heatmaps were drawn with the R-packages gplots [69]. Expression values in the heatmaps with genes correspond to scaled counts, which were $\log_2$-transformed ($\log_2$(equalized counts + 1)). Expression values were further averaged and eventually used to calculate genewise Z-scores.

Gene set enrichment analysis

To test for a polarized auxin or cytokinin response within the syncytial FG, we used the GSEA approach [45, 46]. For GSEA, all genes expressed within the syncytial FG (n = 15'660) are sorted based on their logFC between the two cell halves. This sorted list L is the used to determine whether a set of genes of interest (m) are randomly distributed throughout L or clustered at the top or the bottom (i.e. if they are up- or downregulated). This is achieved by calculating an enrichment score (ES) from a cumulative sum along the list L, which is increased by (n-m)/n whenever a gene is within list of the genes of interest, or decreased by 1 if not. The ES then corresponds to the maximum deviation of the cumulative sum from zero. If the genes of interest were randomly distributed across the sorted list of all genes, the cumulative sum would fluctuate around zero resulting in a small ES. Conversely, a non-random distribution of the genes of interest (for example, accumulation at the bottom of L) would lead to a high ES. An empirical *P*-value can be

calculated by comparing the observed ES to an empirical null-distribution of ES obtained by random sampling of m genes multiple (10'000) times.

Functional characterization

Gene Ontology (GO) annotations, and gene descriptions were retrieved and inferred from ftp.geneontology.org and ftp.arabidopsis.org. GO-term enrichment was calculated with topGO [44] using the "weight" algorithm and Fisher's exact test. For the comparison of an entire transcriptome (i.e. all genes with at least 25 hits in at least half of the replicates of a given tissue or cell type) to the all genes of *Arabidopsis*, P-values were adjusted using Benjamini-Hochberg multiple testing corrections. GO-terms with an FDR < 0.01 were then considered to be significantly enriched in the given transcriptome.

Acknowledgments

We thank Dr. Célia Baroux for help with the confocal microscope. This work was supported by the University of Zurich, and grants from the Swiss National Science Foundation to UG. The study was supported by Life Technologies by contributing, in part, sequencing reagents, which had no influence on the design of the study. PR is supported by the DFG Clusters of Excellence "Future Ocean" and "Inflammation at Interfaces" and the NGFN Network Genomics of Chronic Inflammatory Diseases.

References

[1] Sprunck S, Gro-Hardt R (2011) Nuclear behavior, cell polarity, and cell specification in the female gametophyte. Sexual Plant Reproduction 24: 123-136.

[2] Schmidt A, Wuest SE, Vijverberg K, Baroux C, Kleen D, et al. (2011) Transcriptome analysis of the *Arabidopsis* megaspore mother cell uncovers the importance of RNA helicases for plant germ line development. PLOS Biology 9: e1001155.

[3] Christensen CA, King EJ, Jordan JR, Drews GN (1997) Megagametogenesis in *Arabidopsis* wild type and the Gf mutant. Sexual Plant Reproduction 10: 49-64.

[4] Schneitz K, Grossniklaus U (1998) The molecular and genetic basis of ovule and megagametophyte development. Seminars in Cell and Developmental Biology 9: 227-238.

[5] Schmidt A, Schmid MW, Grossniklaus U (2015) Plant germline formation: molecular insights define common concepts and illustrate developmental flexibility in apomictic and sexual reproduction. Development 142: 229-241.

[6] Huang BQ, Sheridan WF (1994) Female gametophyte development in maize: microtubular organization and embryo sac polarity. The Plant Cell 6: 845-861.

[7] Ebel C, Mariconti L, Gruissem W (2004) Plant retinoblastoma homologues control nuclear proliferation in the female gametophyte. Nature 429: 776-780.

[8] Pagnussat GC, Alandete-Saez M, Bowman JL, Sundaresan V (2009) Auxin-dependent patterning and gamete specification in the *Arabidopsis* female gametophyte. Science 324: 1684-1689.

[9] Ceccato L, Masiero S, Roy DS, Bencivenga S, Roig-Villanova I, et al. (2013) Maternal control of PIN1 is required for female gametophyte development in *Arabidopsis*. PLOS ONE 8: e66148.

[10] Lituiev DS, Krohn NG, Müller B, Jackson D, Hellriegel B, et al. (2013) Theoretical and experimental evidence indicates that there is no detectable auxin gradient in the angiosperm female gametophyte. Development 140: 4544-4553.

[11] Pischke MS, Jones LG, Otsuga D, Fernandez DE, Drews GN, et al. (2002) An *Arabidopsis* histidine kinase is essential for megagametogenesis. PNAS 99: 15800-15805.

[12] Hejátko J, Pernisová M, Eneva T, Palme K, Brzobohatỳ B (2003) The putative sensor histidine kinase CKI1 is involved in female gametophyte development in *Arabidopsis*. Molecular Genetics and Genomics 269: 443-453.

[13] Cheng CY, Mathews DE, Schaller GE, Kieber JJ (2013) Cytokinin-dependent specification of the functional megaspore in the *Arabidopsis* female gametophyte. The Plant Journal 73: 929-940.

[14] Müller B, Sheen J (2007) Advances in cytokinin signaling. Science 318: 68-69.

[15] Gross-Hardt R, Kägi C, Baumann N, Moore JM, Baskar R, et al. (2007) *LACHESIS* restricts gametic cell fate in the female gametophyte of *Arabidopsis*. PLOS Biology 5: e47.

[16] Moll C, Nielsen N, Gross-Hardt R (2008) Mutants with abberant numbers of gametic cells shed new light on old questions. Plant Biology (Stuttgart, Germany) 10: 529-533.

[17] Moll C, von Lyncker L, Zimmermann S, Kägi C, Baumann N, et al. (2008) *CLO/GFA1* and *ATO* are nocel regulators of gametic cell fate in plants. The Plant Journal 56: 913-921.

[18] Shi DQ, Liu J, Xiang YH, Ye D, Sundaresan V, et al. (2005) SLOW WALKER1, essential for gametogenesis in *Arabidopsis*, encodes a WD40 protein involved in 18S ribosomal RNA biogenesis. The Plant Cell 17: 2340-2354.

[19] Liu M, Shi DQ, Yuan L, Liu J, Yang WC (2010) *SLOW WALKER3*, encoding a putative DEAD-box RNA helicase, is essential for female gametogenesis in *Arabidopsis*. Journal of Integrative Plant Biology 52: 817-828.

[20] Pagnussat GC, Yu HJ, Ngo QA, Rajani S, Mayalagu S, et al. (2005) Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. Development 132: 603-614.

[21] Shimizu KK, Okada K (2000) Attractive and repulsive interactions between female and male gametophytes in *Arabidopsis* pollen tube guidance. Development 127: 4511-4518.

[22] Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, et al. (2007) Global analyses of mRNA localization reveals a prominent role in organizing cellular architecture and function. Cell 131: 174-187.

[23] Schönberger J, Hammes UZ, Dresselhaus T (2012) *In vivo* visualization of RNA in plants cells using the λN_{22} system and a GATEWAY-compatible vector series for candidate RNAs. The Plant Journal 71: 173-181.

[24] Honys D, Reňák D, Feciková J, Jedelský PL, Nebesářová J, et al. (2009) Cytoskeleton-associated large RNP complexes in tabacco male gametophyte (EPPs) are associated with ribosomes and are involved in protein synthesis, processing, and localization. Journal of Proteome Research 8: 2015-2031.

[25] Holt CE, Bullock SL (2009) Subcellular mRNA localization in animal cells and why it matters. Science 326: 1212-1216.

[26] Schmid MW, Schmidt A, Klostermeier UC, Barann M, Rosenstiel P, et al. (2012) A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing. PLOS ONE 7: e29685.

[27] Shai JL, Chamberlin MA, Agee A, Caswell ES, Albertsen MC (2013) Transgenic manipulation of plant embryo sacs tracked through cell-type-specific fluorescent markers: cell labeling, cell ablation, and adventitious embryos. Plant Reproduction 26: 125-137.

[28] Kőszegi D, Johnston AJ, Ruttern T, Czihal A, Altschmied L, et al. (2011) Members of the RKD transcription factor family induce and egg cell-like gene expression program. The Plant Journal 66: 890-902.

[29] Rabinger DS, Drews GN (2013) *MYB64* and *MYB119* are required for cellularization and differentiation during female gametogenesis in *Arabidopsis thaliana*. PLOS Genetics 9: e1003783.

[30] Liao Y, Smyth GK, Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research 41: e108.

[31] Schmid MW, Grossniklaus U (2015) Rcount: simple and flexible RNA-Seq read counting. Bioinformatics 31: 436-437.

[32] Gan X, Stegle O, Behr J, Steffen JG, Drewe P, et al. (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477: 419-423.

[33] Tanaka R, Kobayashi K, Masuda T (2011) Tetrapyrrole metabolism in *Arabidopsis thaliana*. The Arabidopsis Book : e0145.

[34] Matias-Hernandez L, Battaglia R, Galbiati F, Rubes M, Eichenberger C, et al. (2010) *VERDANDI* is a direct target of the MADS domain ovule identity complex and affects embryo sac differentiation in *Arabidopsis*. The Plant Cell 22: 1702-1715.

[35] Capron A, Gourgues M, Neica LS, Faure JE, Berger F, et al. (2008) Maternal control of male-gamete delivery in *Arabidopsis* involves a putative GPI-anchored protein encoded by the *LORELEI* gene. The Plant Cell 20: 3038-3049.

[36] Bemer M, Wolters-Arts M, Grossniklaus U, Angenent GC (2008) The MADS domain protein DIANA acts together with AGAMOUS-LIKE80 to specify the central cell in *Arabidopsis* ovules. The Plant Cell 20: 2088-2101.

[37] Steffen JG, Kang IH, Portereiko MF, Lloyd A, Drews GN (2008) AGL61 interacts with AGL80 and is required for central cell development in *Arabidopsis*. Plant Physiology 148: 259-268.

[38] Punwani JA, Rabiger DS, Drews GN (2007) MYB98 positively regulates a battery of synergid-expressed genes encoding filiform apparatus-localized proteins. The Plant Cell 19: 2557-2568.

[39] Diboll AG, Larson DA (1966) An electron micrographic study of the mature megagametophyte in *Zea mays*. Current Opinion in Plant Biology 9: 41-47.

[40] Wuest SE, Vijverberg K, Schmidt A, Weiss M, Gheyselinck J, et al. (2010) *Arabidopsis* female gametophyte gene expression map reveals similarities between plant and animal gametes. Current Biology 20: 1-7.

[41] Schmidt A, Schmid MW, Klostermeier UC, Qi W, Guthörl D, et al. (2014) Apomictic and sexual germline development differ with respect to cell cycle, transcriptional, hormonal and epigenetic regulation. PLOS Genetics 10: e1004476.

[42] Chen C, Farmer AD, Langley RJ, Mudge J, Crow JA, et al. (2010) Meiosis-specific gene discovery in plants: RNA-Seq applied to isolated *Arabidopsis* male meiocytes. BMC Plant Biology 10: 280.

[43] Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11: R25.

[44] Alexa A, Rahnenführer J (2010) topGO: Enrichment analysis for Gene Ontology. R package version 2 18 0.

[45] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics 34: 267-273.

[46] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102: 15545-15550.

[47] Bhargava A, Clabaugh I, To JP, Maxwell BB, Chian YH, et al. (2013) Identification of cytokinin-responsive genes using microarray meta-analysis and rna-seq in *Arabidopsis*. Plant Physiology 162: 272-294.

[48] Rotman N, Durbarry A, Wardle A, Yang WC, Chaboud A, et al. (2005) A novel class of MYB factors controls sperm-cell formation in plants. Current Biology 15: 244-248.

[49] Escobar-Restrepo JM, Huck N, Kessler S, Gagliardini V, Gheyselinck J, et al. (2007) The FERONIA receptor-like kinase mediates male-female interactions during pollen tube reception. Science 317: 606-607.

[50] Kong J, Lau S, Gerd J (2015) Twin plants from supernumerary egg cells in *Arabidopsis*. Current Biology 25: 225-230.

[51] Smyth DR, Bowman JL, Meyerowitz EM (1990) Early flower development in *Arabidopsis*. The Plant Cell 2: 755-767.

[52] De Rybel B, van den Berg W, Lokerse AS, Liao CY, van Mourik H, et al. (2011) A versatile set of Ligation-Independent Cloning vectors for functional studies in plants. Plant Physiology 156: 1292-1299.

[53] Logemann E, Birkenbihl RP, Ülker B, Somssich IE (2006) An improved method for preparing *Agrobacterium* cells that simplifies the *Arabidopsis* transformation protocol. Plant Methods 2: 16.

[54] Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, et al. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Research 20: 45-58.

[55] Lister R, O'Malley RC, Tonti-Filippini J, Gregory BG, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133: 523-536.

[56] Nodine MD, Bartel DP (2010) MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. Genes & Development 24: 2678-2692.

[57] Autran D, Baroux C, Raissig MT, Lenormand T, Wittig M, et al. (2011) Maternal epigenetic pathways control parental contributions to *Arabidopsis* early embryogenesis. Cell 145: 707-719.

[58] Abeel T, Van Parys T, Saeys Y, Galagan J, de Peer Y V (2012) GenomeView: a next-generation genome browser. Nucleic Acids Research 40: e12.

[59] Curtis MD, Grossniklaus U (2003) A gateway cloning vector set for high-throughput functional analysis of genes in planta. Plant Physiology 133: 462-469.

[60] Huck N, Moore JM, Federer M, Grossniklaus U (2003) The *Arabidopsis* mutant *feronia* disrupts the female gametophytic control of pollen tube reception. Development 130: 2149-2159.

[61] Gehring M, Missirian S V Henikoff (2011) Genomic analysis of parent-of-origin allelic expression in *Arabidopsis thaliana* seeds. PLOS ONE 6: e23687.

[62] Torti S, Fornara F, Vincent C, Andrés F, Nordström K, et al. (2012) Analysis of the *Arabidopsis* shoot meristem transcriptome during floral transition identifies distinct regulatory patterns and a leucine-rich repeat protein that promotes flowering. The Plant Cell 24: 444-462.

[63] Loraine AE, McCormick S, Estrada A, Patel K, Qin P (2013) RNA-Seq of *Arabidopsis* pollen uncovers novel transcription and alternative splicing. Plant Physiology 162: 1092-1109.

[64] Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, et al. (2013) Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods 10: 1093-1095.

[65] Wang H, Chung PJ, Liu J, Jang IC, Kean MJ, et al. (2014) Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in *Arabidopsis*. Genome Research 24: 3.

[66] Zhu Y, Rowley MJ, Böhmdorfer G, Wierzbicki AT (2013) A SWI/SNF chromatin-remodeling complex acts in noncoding RNA-mediated transcriptional silencing. Molecular Cell 49: 298-309.

[67] Ausin I, Greenberg MV, Simanshu DK, Hale CJ, Vashisht AA, et al. (2012) INVOLVED IN DE NOVO 2-containing complex involved in RNA-directed DNA methylation in *Arabidopsis*. PNAS 109: 8374-8381.

[68] Stroud H, Hale CJ, Feng S, Caro E, Jacob Y, et al. (2012) DNA methyltransferases are required to induce heterochromatic re-replication in *Arabidopsis*. PLOS Genetics 8: e1002808.

[69] Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, et al. (2015) gplots: Various R Programming Tools for Plotting Data. R package version 2 16 0.

[70] Schmidt A, Schmid MW, Grossniklaus U (2012) Analysis of plant germline development by high-throughput RNA profiling: technical advances and new insights. The Plant Journal 70: 18-29.

[71] Colombo M, Masiero S, Vanzulli S, Lardelli P, Kater MM, et al. (2008) *AGL23*, a type I MADS-box gene that controls female gametophyte and embryo development in *Arabidopsis*. The Plant Journal 54: 1037-1048.

[72] de Folter S, Immink RG, Kieffer M, Parenicová L, Henz SR, et al. (2005) Comprehensive interaction map of the *Arabidopsis* MADS box transcription factors. The Plant Cell 17: 1424-1433.

[73] Portereiko MF, Lloyd A, Steffen JG, Punwani JA, Otsuga D, et al. (2006) AGL80 is required for central cell and endosperm development in *Arabidopsis*. The Plant Cell 18: 1862-1872.

[74] Acosta-Garcia G, Vielle-Calzada JP (2004) A classical arabinogalactan protein is essential for the initiation of female gametogenesis in *Arabidopsis*. The Plant Cell 16: 2614-2628.

[75] Capron A, Serralbo O, Fülöp K, Frugier F, Parmentier Y, et al. (2003) The *Arabidopsis* anaphase-promoting complex or cyclosome: molecular and genetic characterization of the APC2 subunit. The Plant Cell 15: 2370-2382.

[76] Huanca-Mamani W, Garcia-Aguilar M, León-Martínez G, Grossniklaus U, Vielle-Calzada JP (2005) CHR11, a chromatin-remodeling factor essential for nuclear proliferation during female gametogenesis in *Arabidopsis thaliana*. PNAS 102: 17231-17236.

[77] Choi Y, Gehring M, Johnson L, Hannon M, Harada JJ, et al. (2002) DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in *Arabidopsis*. Cell 110: 33-42.

[78] Chaudhury AM, Ming L, Miller C, Craig S, Dennis ES, et al. (1997) Fertilization-independent seed development in *Arabidopsis thaliana*. PNAS 94: 4223-4228.

[79] Sung AO, Johnson A, Smertenko A, Rahman D, Soon KP, et al. (2005) A divergent cellular role for the FUSED kinase family in the plant-specific cytokinetic phragmoplast. Current Biology 15: 2107-2111.

[80] Alandete-Saez M, Ron M, McCormick S (2008) *GEX3*, expressed in the male gametophyte and in the egg cell of *Arabidopsis thaliana*, is essential for micropylar pollen tube guidance and plays a role during early embryogenesis. Molecular Plant 1: 586-598.

[81] Christensen CA, Gorsich SW, Brown RH, Jones LG, Brown J, et al. (2002) Mitochondrial GFA2 is required for synergid cell death in *Arabidopsis*. The Plant Cell 14: 2215-2232.

[82] Niewiadomski P, Knappe S, Geimer S, Fischer K, Schulz B, et al. (2005) The *Arabidopsis* plastidic glucose 6-phosphate/phosphate translocator GPT1 is essential for pollen maturation and embryo sac development. The Plant Cell 17: 760-775.

[83] Cigliano RA, Cremona G, Paparo R, Termolino P, Perrella G, et al. (2013) Histone deacetylase AtHDA7 is required for female gametophyte and embryo development in *Arabidopsis*. Plant Physiology 163: 431-440.

[84] Guitton AE, Berger F (2005) Loss of function of MULTICOPY SUPPRESSOR OF IRA 1 produces nonviable parthenogenetic embryos in *Arabidopsis*. Current Biology 15: 750-754.

[85] Tanaka H, Ishikawa M, Kitamura S, Takahashi Y, Soyano T, et al. (2004) The *AtNack1/HINKEL* and *STUD/TETRASPORE/AtNACK2* genes, which encode functionally redundant kinesins, are essential for cytokinesis in *Arabidopsis*. Genes to Cells 9: 1199-1211.

[86] Kwee HS, Sundaresan V (2003) The *NOMEGA* gene required for female gametophyte development encodes the putative APC6/CDC16 component of the anaphase promoting complex in *Arabidopsis*. The Plant Journal 36: 853-866.

[87] Pagnussat GC, Yu HJ, Sundaresan V (2007) Cell-fate switch of synergid to egg cell in *Arabidopsis eostre* mutant embryo sacs arises from misexpression of the bel1-like homeodomain gene *BLH1*. The Plant Cell 19: 3578-3592.

[88] Gallois JL, Guyon-Debast A, Lécureuil A, Vezon D, Carpentier V, et al. (2009) The *Arabidopsis* proteasome RPT5 subunits are essential for gametophyte development and show accession-dependent redundancy. The Plant Cell 21: 442-459.

[89] Huang CK, Huang LF, Huang JJ, Wu SJ, Yeh CH, et al. (2010) A DEAD-box protein, AtRH36, is essential for female gametophyte development and is involved in rRNA biogenesis in *Arabidopsis*. Plant Cell Physiology 51: 694-706.

[90] Kägi C, Baumann N, Nielsen N, Stierhof YD, Gross-Hardt R (2010) The gametic central cell of *Arabidopsis* determines the lifespan of adjacent accessory cells. PNAS 107: 22350-22355.

[91] Pastuglia M, Azimzadeh J, Goussot M, Camilleri C, Belcram K, et al. (2006) γ-tubulin is essential for microtubule organization and development in *Arabidopsis*. The Plant Cell 18: 1412-1425.

Figures

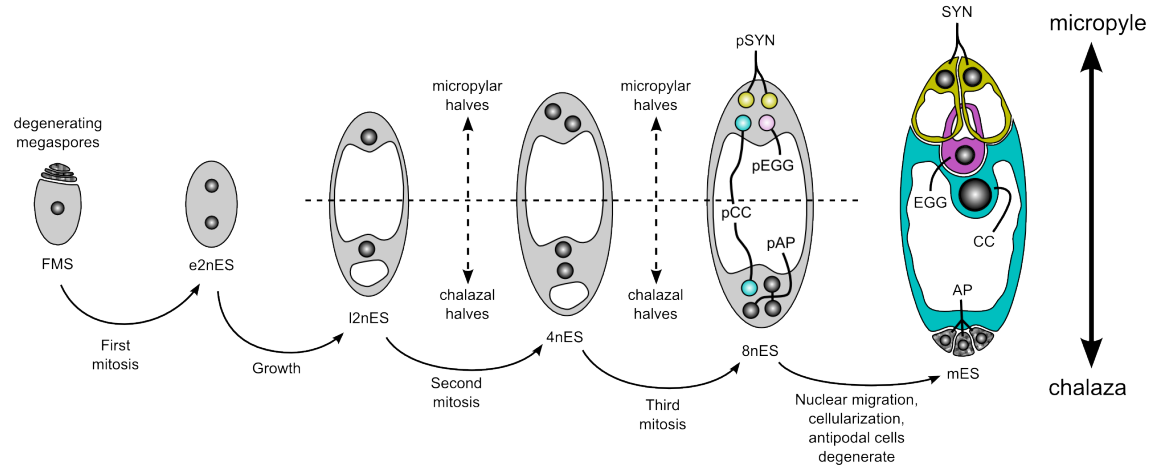


Figure 1. Schematic representation of female gametogenesis in *Arabidopsis thaliana*. Female gametogenesis starts with a sporophytic cell differentiating into a megaspore mother cells (MMC). The MMC then undergoes meiosis to form a tetrad of four haploid spores, three of which degenerate. The remaining functional megaspore (FMS) subsequently undergoes three rounds of mitosis in a syncytium. Nuclear migration and cellularization then eventually lead to the formation of a mature female gametophyte containing seven cells: the central cell, the egg cell, two synergids, and three antipodals. Abbreviations: e2nES/l2nES, early/late two-nucleate embryo sac (FG2/FG3); 4nES, four-nucleate embryo sac (FG4); 8nES, eight-nucleate embryo sac (FG5/FG6); (p)AP, (precursor of) antipodal cell; (p)EGG, (precursor of) egg cell; (p)SYN, (precursor of) synergid cell; (p)CC, (precursor of) central cell; mES, mature embryo sac (FG7). Modified after [70].

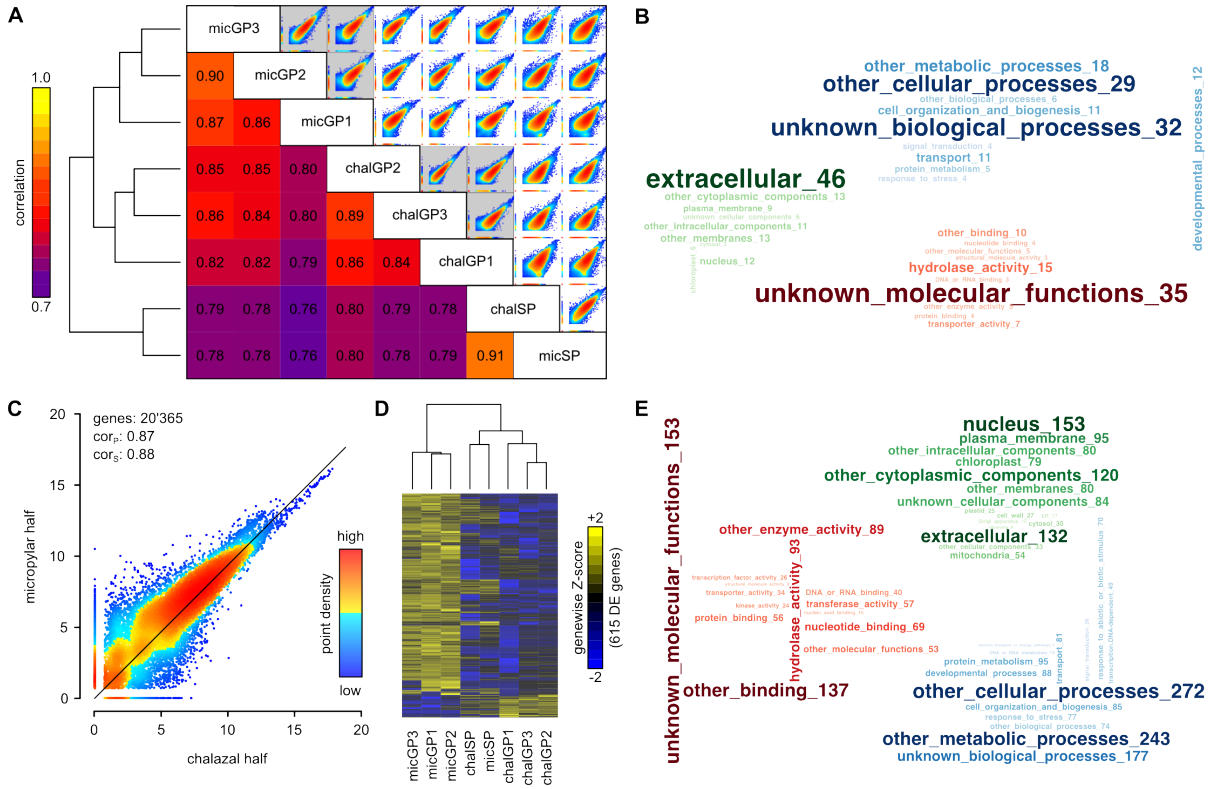


Figure 2. The transcriptome of the micropylar and chalazal poles of the syncytial female gametophyte. (A) Sample correlation and pairwise comparison of gene expression values illustrating the reproducibility of the gametophytic triplicates and the higher similarity between the gametophytic samples compared to the surrounding sporophytic samples. Expression values correspond to unscaled raw data, which were $\log_2$-transformed ($\log_2$(raw counts + 1)). Samples were clustered using Pearson correlation and hierarchical agglomerative clustering (complete linkage). Abbreviations: micGPx/chalGPx, micropylar/chalazal gametophyte; micSP/chalSP, micropylar/chalazal sporophyte. (B) GO-Slim term annotation of the 101 genes specifically expressed in the syncytial FG represented as word clouds. The size of a word is proportional to its occurence in the data set. Occurences are added behind the term. Blue, green, and red for terms belonging to the GO-domains "biological process", "cellular component", and "molecular function", respectively. (C) Comparison of average gene expression values illustrating the high overall-similarity between the two cell halves with some genes exhibiting differential expression. Expression values correspond to average scaled counts, which were $\log_2$-transformed (mean($\log_2$(equalized counts + 1))). Counts were equalized with edgeR [43]. Only genes with at least five reads over all samples are displayed. (D) Expression values of genes differentially expressed within the FG. Out of 20'365 genes tested, 615 genes were found to be polarized within the syncytial FG (FDR < 0.05, absolute log fold-change (logFC) > 2, 554/61 genes in the micropylar and chalazal halves, respectively). Surrounding sporophytic samples (micSP, chalSP) are shown as well. Expression values correspond to genewise Z-scores of scaled, $\log_2$-transformed count data. Samples/genes were clustered using euclidean distance and hierarchical agglomerative clustering (complete linkage). (E) As in (B), but for the 615 polarized genes.

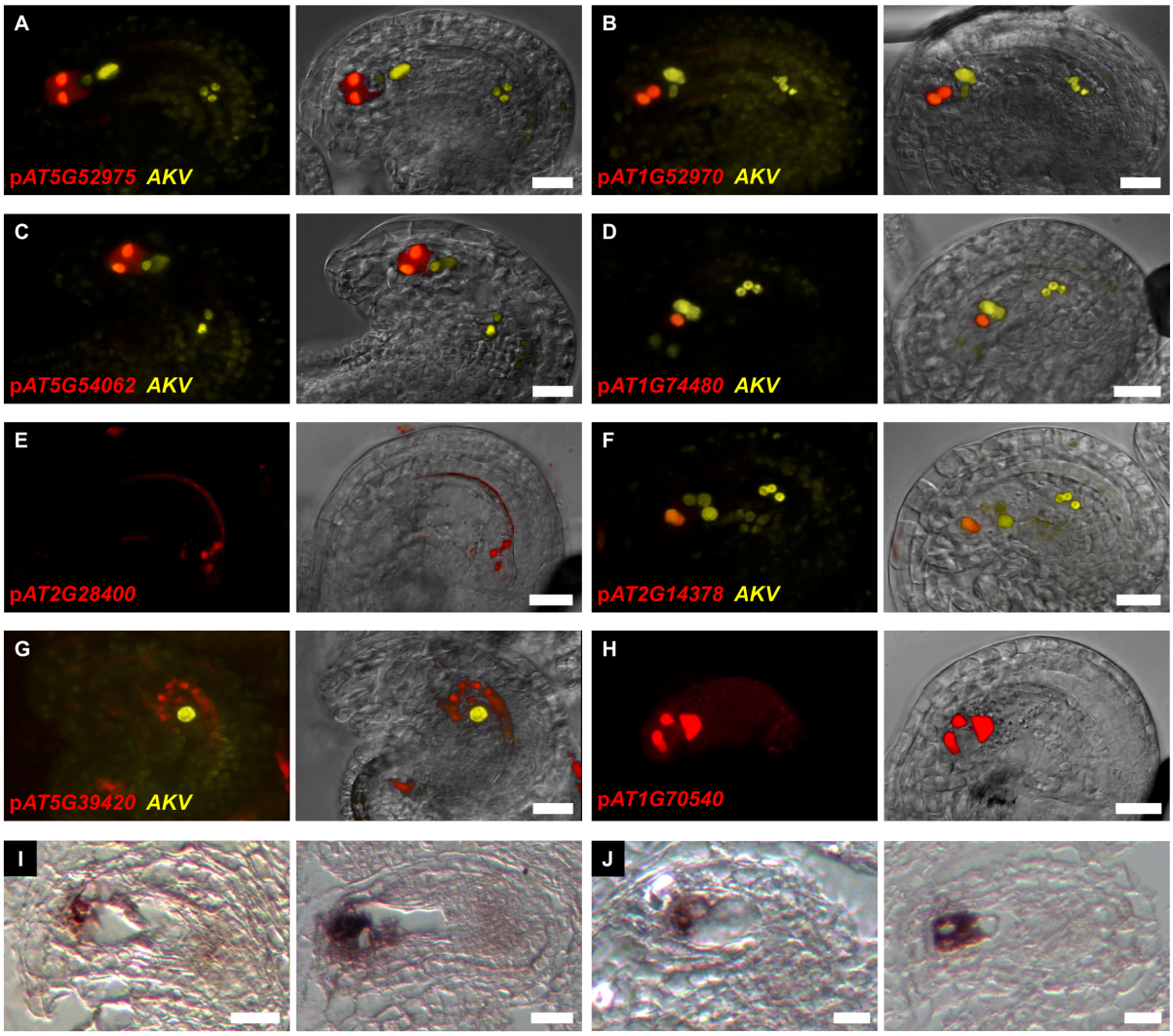


Figure 3. Data validation by reporter constructs and RNA *in situ* hybridizations. Data validation for polarized genes with RNA *in situ* hybridization (I,J) and reporter constructs (nuclear-localized red fluorescent protein) in a *AKV* marker background [48] expressing a nuclear-localized yellow fluorescent protein in all cells of the female gametophyte starting from the FG1 stage on (A-H, except E and H, which were in Landsberg *erecta*). (A-H) Several reporters were found to be active in the synergids (A,B,C,F). Other reporters displayed a signal in the egg cell (D), in the micropylar nuclei at the FG5 stage (not shown) and the micropylar cells of the mature female gametophyte (H), in the micropylar sporophyte at early stages of FG development and the vasculature (G), and the cells adjacent to the chalazal half of the syncytial female gametophyte (E). (I) *AT2G38750* is expressed in the micropylar half of the syncytial FG starting from the FG4 stage on (left) and later in the synergids and the egg cell (right). (J) *AT4G35165* is expressed in the micropylar half of the syncytial FG starting from the FG5 stage on (left) and later in the synergids (right). (I,J) For both RNA *in situ* hybridization probes, the signals could only be observed with the antisense probe. No signals were observed using sense probes (not shown). (A-J) Scale bar equals to 25 μm. See as well Table 2.

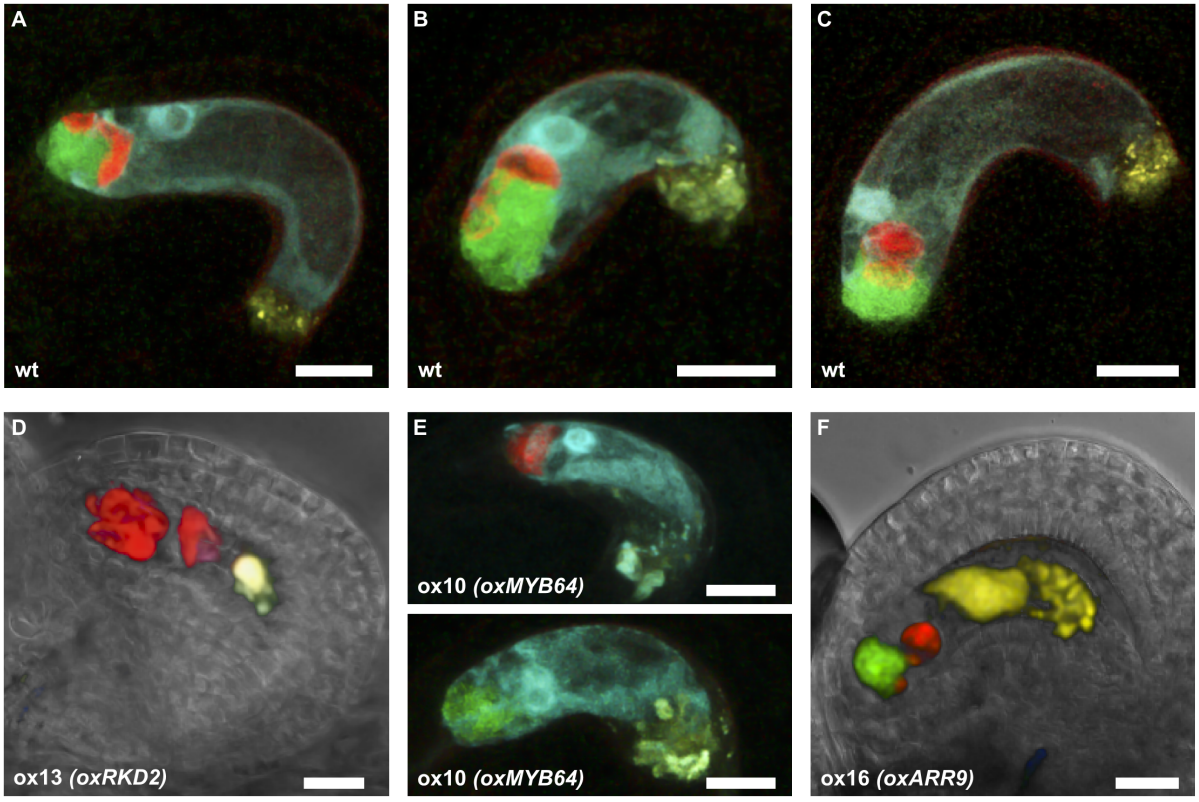


Figure 4. Screening candidate genes for alterations of cell fate. Constructs for uniform overexpression of candidate genes were transformed into a plant line marking all four cell types of the mature FG [27] and screened for alterations in cell fate. In addition, high rates of ovule abortion could qualitatively be identified. (A-C) Normal female gametophytes with two synergids (green), one egg cell (red), one central cell (blue), and three antipodal cells (yellow, marked endothelium as well). (D-F) Ovules carrying a construct for uniform expression of a candidate genes during the entire female gametogenesis. Only candidates which were characterized in greater detail are shown. (D) *RKD2*, central cell and synergids express the egg cell marker (red). (E) *MYB64*, either synergids (top) or egg cell (bottom) are sometimes not specified. Unclear if due to premature cellularization. High rate of ovule abortion (not shown). (F) *ARR9*, central cell frequently express the antipodal marker (yellow) and sometimes appear to move towards the chalazal pole of the female gametophyte. High rate of ovule abortion (not shown). Note that this image was aquired differently than all others in the panel (see Material and Methods, screen first batch). (A-F) Scale bar equals to 25 μm.

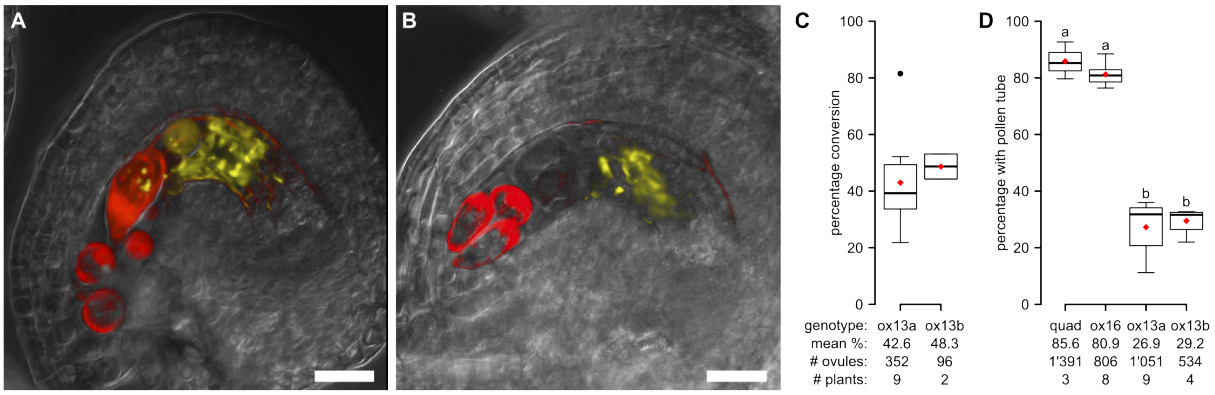


Figure 5. Uniform overexpression of *RKD2* during female gametogenesis. (A,B) Ovules expressing the egg cell inducer *RKD2* under the control elements of the *AKV* reporter [48]. Aside the induction of the marker expression, cell shape and position were generally altered as well (A). The cell positions were rarely similar as in wild-type (B). Note, however, that in this case, the synergids displayed an altered polarity similar to the egg cell with the vacuole located at the micropylar side of the cell and the nucleus on the opposite side. See Figure 4(A,B,C) for examples of wild-type FGs. (A,B) Only RFP and YFP channel were recorded (no signal was present in the other two). (C) Quantification of conversion of the cells of the mature female gametophyte into cells expressing the egg cell marker in two independent lines. The numbers give the percentages of ovules displaying two or more cells expressing the marker. In the majority of the ovules, synergids and central cell expressed the egg cell marker. Antipodal cells frequently still expressed the antipodal marker. Note, however, that the antipodal signal was difficult to unambiguously assign to the cells of the female gametophyte, given that the surrounding endothelium showed expression of the marker as well. (D) Quantification of pollen tube reception with aniline blue staining of callose in the pollen tubes in two independent lines. The numbers refer to the percentage of ovules successfully attracting a pollen tube one day after pollination with pollen from the quadruple marker line. The line "ox16" expressing *ARR9* under the control elements of the *AKV* reporter [48] served as a control for the presence of an additional transgene. Letters above the boxplots indicate if two lines were significantly different from each other in a pairwise comparison (shared letter means no significant difference). (C,D) Plants used for quantification were offspring from two independent T1 plants (ox13a and ox13b). Means are shown as red diamonds.

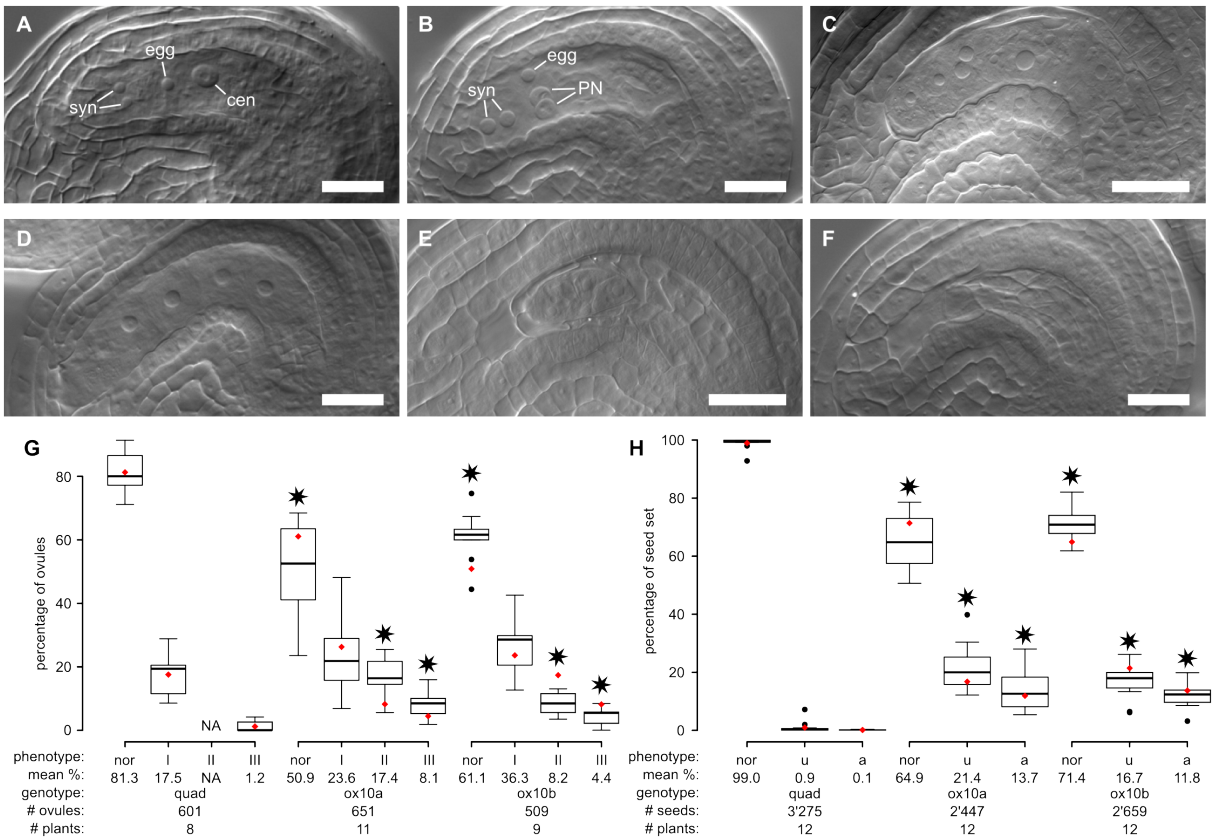


Figure 6. Uniform overexpression of *MYB64* during female gametogenesis. (A) A normal female gametophyte (quad marker). The two synerid nuclei (syn), the egg cell nucleus (egg), and the large central cell nucleus (cen) are well visible. (B-F) Plants carrying a construct for uniform overexpression of *MYB64* displayed a variety of abnormally shaped female gametophytes. We classified them into three groups: (I, in B) Individual cells are still recognizable, eventually slightly misplaced, and polar nuclei (PN) remain unfused. (II, in C-E) Cell types can not be distinguished anymore, because nuclei/cells are sometimes equally sized and/or positioning of the cells is distorted. (III, in F) Complete degradation or absence of the female gametophyte. (G) Quantification of the phenotypes illustrated in (A-F) in two independent lines. Note that the first (unfused PN) and the last group (FG completely missing) can also be observed in wild-type ovules. Asterisks mark significant differences compared to the quadruple marker background (or significant difference from zero for the class II phenotypes). (H) Quantification of seed set, i.e., the number of normal (nor), unfertilized (u), and aborted (a) ovules, in two independent lines. Unfertilized/pre-fertilization aborted ovules appeared white, whereas post-fertilization aborted ovules were markedly bigger and brown. Asterisks mark significant differences compared to the quadruple marker background. (A-F) Scale bars correspond to 20 μm. (G,H) Note that the wild-type phenotype and seed counts are the same as in Figure 7D,E. Plants used for quantification were offspring from two independent T1 plants (ox10a and ox10b). Significance level for the asterisk was set to P-value < 0.05. Means are shown as red diamonds.

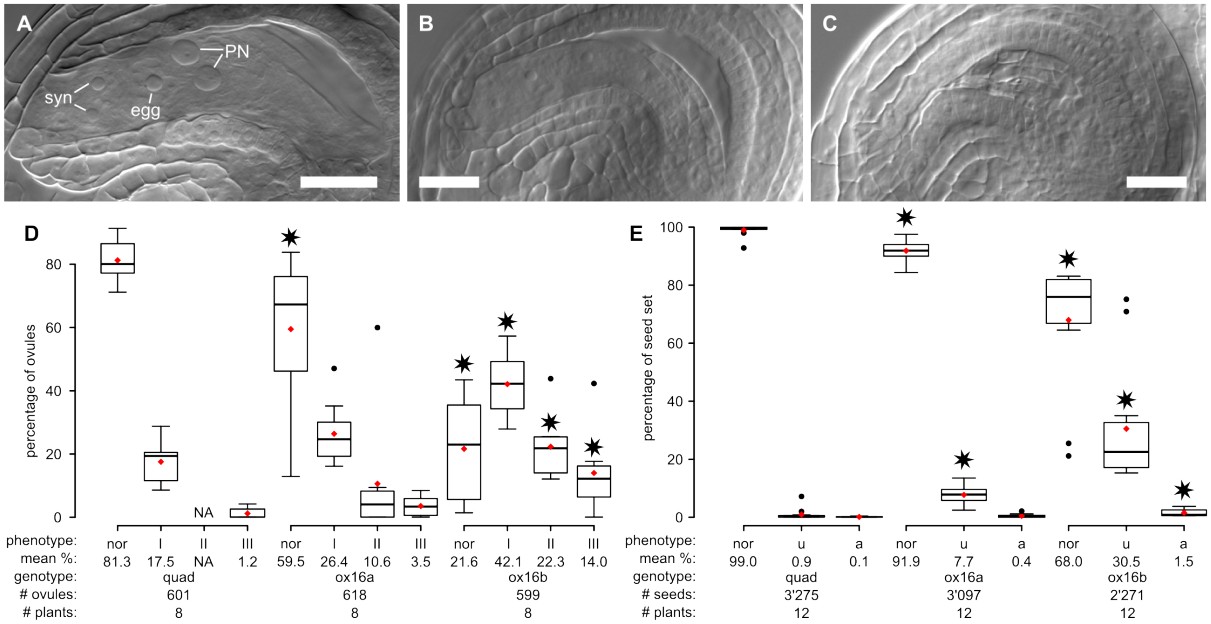


Figure 7. Uniform overexpression of *ARR9* during female gametogenesis. For a normal female gametophyte (quad marker), see Figure 6A. (A-C) Plants carrying a construct for uniform overexpression of *ARR9* displayed three distinguishable phenotypes: (I, in A) Individual cells were still recognizable but polar nuclei (PN) remain unfused. (II, in B) Certain cell/nuclei types, i.e. mainly the central cell nucleus and the antipodal cell nuclei, were not distinguishable anymore and were frequently equally-sized. Antipodal cells were sometimes shifted towards the central cell/polar nuclei (not shown). (III, in C) Complete degradation or absence of the female gametophyte. (D) Quantification of the phenotypes illustrated in (A-C) in two independent lines. Note that the first (unfused PN) and the last group (FG completely missing) can also be observed in wild-type ovules. Asterisks mark significant differences compared to the quadruple marker background (or significant difference from zero for the class II phenotypes). (H) Quantification of seed set, i.e., the number of normal (nor), unfertilized (u), and aborted (a) ovules, in two independent lines. Unfertilized/pre-fertilization aborted ovules appeared white, whereas post-fertilization aborted ovules were markedly bigger and brown. Asterisks mark significant differences compared to the quadruple marker background. (A-C) Scale bars correspond to 20 μm. (D,E) Note that the wild-type phenotype and seed counts are the same as in Figure 6G,H. Plants used for quantification were offspring from two independent T1 plants (ox16a and ox16b). Significance level for the asterisk was set to P-value < 0.05. Means are shown as red diamonds.

Tables

Table 1. Classification of alignments. The table summarizes the number of hits found in a certain genomic region. Percentages are given in parentheses (percent aligned or percent of aligned).

sample	reads total	thereof aligned	exons and splice junctions	introns	intergenic
micSP	43'479'028	18'013'618 (41.4)	16'096'442 (89.4)	827'932.6 (4.6)	1'089'243 (6.0)
micGP1	46'569'778	19'327'097 (41.5)	17'086'395 (88.4)	958'391.6 (5.0)	1'282'311 (6.6)
micGP2	49'787'731	17'242'792 (34.6)	15'433'403 (89.5)	753'810.8 (4.4)	1'055'578 (6.1)
micGP3	50'592'110	18'366'668 (36.3)	16'426'607 (89.4)	774'488.8 (4.2)	1'165'572 (6.3)
chalSP	46'966'876	18'457'076 (39.3)	16'542'618 (89.6)	682'524.3 (3.7)	1'231'934 (6.7)
chalGP1	36'822'330	21'172'932 (57.5)	18'798'223 (88.8)	1'212'532 (5.7)	1'162'177 (5.5)
chalGP2	46'235'185	7'996'882 (17.3)	7'185'395 (89.9)	406'607.7 (5.1)	404'879.5 (5.1)
chalGP3	48'727'765	6'304'992 (12.9)	5'655'384 (89.7)	291'048.8 (4.6)	358'559.1 (5.7)

Table 2. Data validation by reporter constructs and RNA *in situ* hybridizations. Top: Putative promoters of polarized genes were cloned in front of a reporter gene. Out of fourteen candidates, eight exhibited a detectable signal. Bottom: Results from RNA *in situ* hybridization experiments. The table gives a summary of the localization of the signals. The expression pattern of the *AT1G70540* reporter was variable. Starting from FG4 stage on, it was mostly only detectable in the micropylar nuclei of the syncytial FG and the egg apparatus and the central cell of the mature FG. However, the signal could sometimes also be detected within all nuclei of the FG.

AGI	gene	up in	# lines	localization
AT5G52975	*DUF1278*	mic	5	synergids
AT1G52970	*DD11*	mic	3	synergids
AT5G54062	no name	mic	3	synergids
AT1G74480	*RKD2*	mic	7	egg
AT2G28400	*DUF584*	chal	7	cells adjacent to chalazal gametophyte, weak
AT2G14378	*ECA1*-related	mic	10	synergids
AT5G39420	*CDC2C*	mic	4	vasculature, micropylar sporophyte during early FG stages
AT1G70540	*EDA24*	mic	2	micropylar nuclei FG5/egg apparatus/central cell
AT2G38750	*ATANN4*	mic	-	micropylar half from FG4 stage on/egg/synergids
AT4G35165	*DUF1278*	mic	-	micropylar half at FG5 stage/synergids

Table 3. Genes with a potential role in female gametophyte specification. Genomic sequences without the 5' and 3' untranslated regions of 34 polarized genes were cloned into a vector driving uniform expression within the whole female gametophyte starting from the functional megaspore (FG1) stage on throughout gametogenesis and in the mature FG. Seven candidates displayed alterations of cell fate or apparently high percentage of aborted ovules (qualitative observation from the screen with the confocal microscope) within at least three transformants. T1 a/t: number of T1 plants with alterations/in total.

ID	AGI	gene	up in	T1 a/t	observation
ox8	*AT5G15960*	*KIN1*	chal	4/10	egg/synergids sometimes not specified
ox10	*AT5G11050*	*MYB64*	mic	6/13	abnormal cell shape, aborted ovules
ox11	*AT5G58850*	*MYB119*	mic	3/9	abnormal cell shape, aborted ovules
ox13	*AT1G74480*	*RKD2*	mic	4/8	multiple egg cells
ox16	*AT3G57040*	*ARR9*	chal	4/12	central cell sometimes expresses AP marker, aborted ovules
ox34	*AT2G37010*	*NAP12*	mic	3/4	aborted ovules
ox37	*AT4G38070*	bHLH-TF	chal	4/20	rarely shrunken central cell and unspecified egg, aborted ovules

Supporting Information

S1 Figure

Sample isolation using LAM. Isolation of the four-nucleate embryo sac samples. Upper row: micropylar half of the embryo sac (mG) and the corresponding sporophytic control (mS). Lower row: chalazal half of the embyro sac (cG) and the corresponding sporophytic control (cS). A: 7 μm thin section through an ovule carrying a four-nucleate embryo sac before LAM. Only the two micropylar nuclei are visible. B: the micropylar half of the embryo sac has been cut and collected. C: the micropylar sporophytic control after cutting. D: The micropylar sporophytic control has been removed using the isolation cap as described [40]. E: The two chalazal nuclei of the gametophyte are visible. F and G: the chalazal half of the gametophyte has been cut and isolated. H and I: the sporophytic control has been cut and isolated.

S2 Figure

Similarity between tissue and cell types. Heatmap illustrating the similarity between different tissue and cell types (Pearson correlation coefficients). Tissue and cell types were clustered based on their average gene expression profiles using Pearson correlation and hierarchical agglomerative clustering.

S3 Figure

Comparison of alignment statistics. Heatmap illustrating the similarity between different samples in terms of their basic alignment statistics. Each alignment of a read can be characterized by several aspects. First, a read may align uniquely or multiple times to the genome. The alignment can map either to an intergenic region or to a gene. In both cases, the alignment may be gapped. Hits mapping to genes can be separated into intronic (i.e. not matching to known exons/splice junctions), exonic, or covering a splice junction. The latter two can in addition be separated into ambiguous (mapping to more than one gene) or unambiguous (mapping to only one gene). Samples were clustered using Pearson correlation and hierarchical agglomerative clustering. The dissimilarity between some of the samples was due to differences in the number of reads with multiple alignments and reads originating from intergenic regions (see S3 Table).

S4 Figure

Genes significantly enriched in the female gametophyte. Expression values of genes preferentially expressed in the female gametophyte are summarized in a heatmap (blue/yellow: low/high expression values). Genewise Z-scores were calculated with the average of scaled, $\log_2$-transformed count data (mean($\log_2$(equalized counts + 1))). Cell/tissue

types and genes were clustered using euclidean distance and hierarchical agglomerative clustering.

S5 Figure

Expression pattern of small nucleolar RNAs (snoRNAs). Expression values of 68/71 snoRNA genes are summarized in a heatmap (blue/yellow: low/high expression values, the remaining three genes were not sequenced in any library). Genewise Z-scores were calculated with the average of scaled, $\log_2$-transformed count data (mean($\log_2$(equalized counts + 1))). Cell/tissue types and genes were clustered using euclidean distance and hierarchical agglomerative clustering.

S6 Figure

Gene set enrichment analysis of genes responsive to auxin or cytokinin stimulus. Test for enrichment of genes responsive to auxin (A) or cytokinin (B) in one of the two halves of the syncytial FG. Line graph on top shows the cumulative sum calculated along the list of genes sorted according to their logFC between the two halves of the syncytial FG (see Material and Methods for details). For a random set of genes, the cumulative sum would fluctuate around zero and the maximum deviation from zero (i.e. the enrichment score ES) would be small. Enrichment of the genes in either half of the FG would be visible as a cluster of genes on the top or the bottom of the sorted genes and a corresponding high ES. An empirical P-value can be calculated by comparing the observed ES to an empirical null-distribution of ES obtained through random sampling. The distribution of the 20 tested auxin-responsive genes does not differ significantly from a random distribution. In contrast, the 73 tested cytokinin-responsive genes show a highly significant enrichment in the chalazal half of the syncytial FG (P-value < 0.0001). Red lines mark the genes in the gene set of interest.

S1 Table

RNA-Seq data used in this study. Data used for RNA-Seq analysis. I for Illumina, S for SOLiD, SE and PE for single- and paired-end, respectively. 4nES: four-nucleate embryo sac, SAM: shoot apical meristem, SD/LD: short/long day conditions. All PE samples were treated as SE (reverse reads were removed).

S2 Table

Expression data of all samples processed in this study. The zip-file contains a tab-separated table with the number of hits per gene for all samples listed in S1 Table.

S3 Table

Alignment statistics for all RNA-Seq samples. The table contains basic alignment statistics for all RNA-Seq samples used in this study. The data were visualized in S3 Figure.

S4 Table

GO terms specifically enriched in the developing female gametophyte. The table contains all genes expressed within the developing female gametophyte, which belong to one of the GO-terms found to be specifically enriched only in the transcriptome of the developing female gametophyte (GO:0051302, GO:0009567, and GO:0033013).

S5 Table

Genes involved in cell fate determination and polarity of the embryo sac. Genes playing a role in polarity, determination of cell fate, and gametophyte development in general were inferred from mutants described in the literature [7, 11–13, 15–21, 29, 34, 36–38, 50, 71–91] and tested for their expression pattern in the developing female gametophyte. Genes likely important at very early stages (e.g., meiotic genes) or for a specific function of a mature cell (e.g., genes involved in pollen tube attraction) were not included. Candidates (EMBRYO SAC DEVELOPMENTAL ARREST) from [20] were only taken if they were characterized in greater detail (MATERNAL EFFECT EMBRYO ARREST and UNFERTILIZED EMBRYO SAC were not included as they are likely to act relatively late during FG development). Note that the citation numbering in the table is different from the one in the main text.

S6 Table

Genes highly enriched in the developing female gametophyte. A table with the 101 genes identified to highly enriched within the developing female gametophyte compared to all other samples listed in S1 Table containing gene type, short/long gene description, geneFamily, GO-term (biological process, molecular function, and cellular components) annotations, PFAM annotations, and InterPro (family, domain, conserved site, binding site, active site, repeat, post-translational modification) annotations.

S7 Table

GO terms significantly enriched in the female gametophyte. Significantly enriched GO terms in the set of genes specific to the female gametophyte (including the synergids, egg, and central cell samples) compared to all genes expressed within the tested samples. GO-term enrichment was calculated with topGO [44] using the "weight"

algorithm and Fisher's exact test. Obs/exp: number of genes observed/expected within this term.

S8 Table

Genes differentially expressed within the developing female gametophyte (polarized genes). A table with the 615 genes identified to be differentially expressed between the two cell halves containing average expression values, standard deviations, FDR-adjusted *P*-values, gene type, short/long gene description, geneFamily, GO-term (biological process, molecular function, and cellular components) annotations, PFAM annotations, and InterPro (family, domain, conserved site, binding site, active site, repeat, post-translational modification) annotations.

S9 Table

GO-term enrichment within the list of polarized genes. Significantly enriched GO terms in the list of polarized genes compared to all genes having at least five reads within one of the six gametophytic samples. GO-term enrichment was calculated with topGO [44] using the "weight" algorithm and Fisher's exact test. Obs/exp: number of genes observed/expected within this term.

S1 File

Library quality controls. The file contains the results from the cDNA library control experiments (size distribution of fragments and approximate concentration of selected genes).

S2 File

Primer sequences. The file contains all primer sequences used in this study.

Sample isolation using LAM

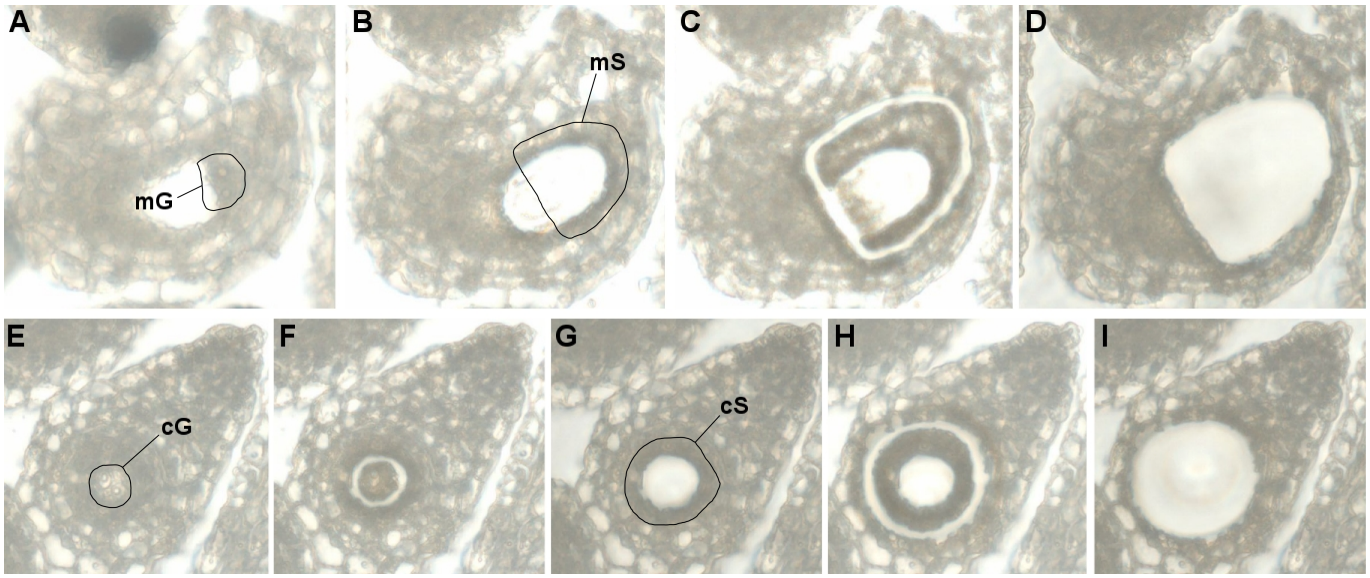


Figure S1: Isolation of the four-nucleate embryo sac samples. Upper row: micropylar half of the embryo sac (mG) and the corresponding sporophytic control (mS). Lower row: chalazal half of the embyro sac (cG) and the corresponding sporophytic control (cS). A: 7 μm thin section through an ovule carrying a four-nucleate embryo sac before LAM. Only the two micropylar nuclei are visible. B: the micropylar half of the embryo sac has been cut and collected. C: the micropylar sporophytic control after cutting. D: The micropylar sporophytic control has been removed using the MMI isolation cap. E: The two chalazal nuclei of the gametophyte are visible. F and G: the chalazal half of the gametophyte has been cut and isolated. H and I: the sporophytic control has been cut and isolated.

Similarity between tissue and cell types

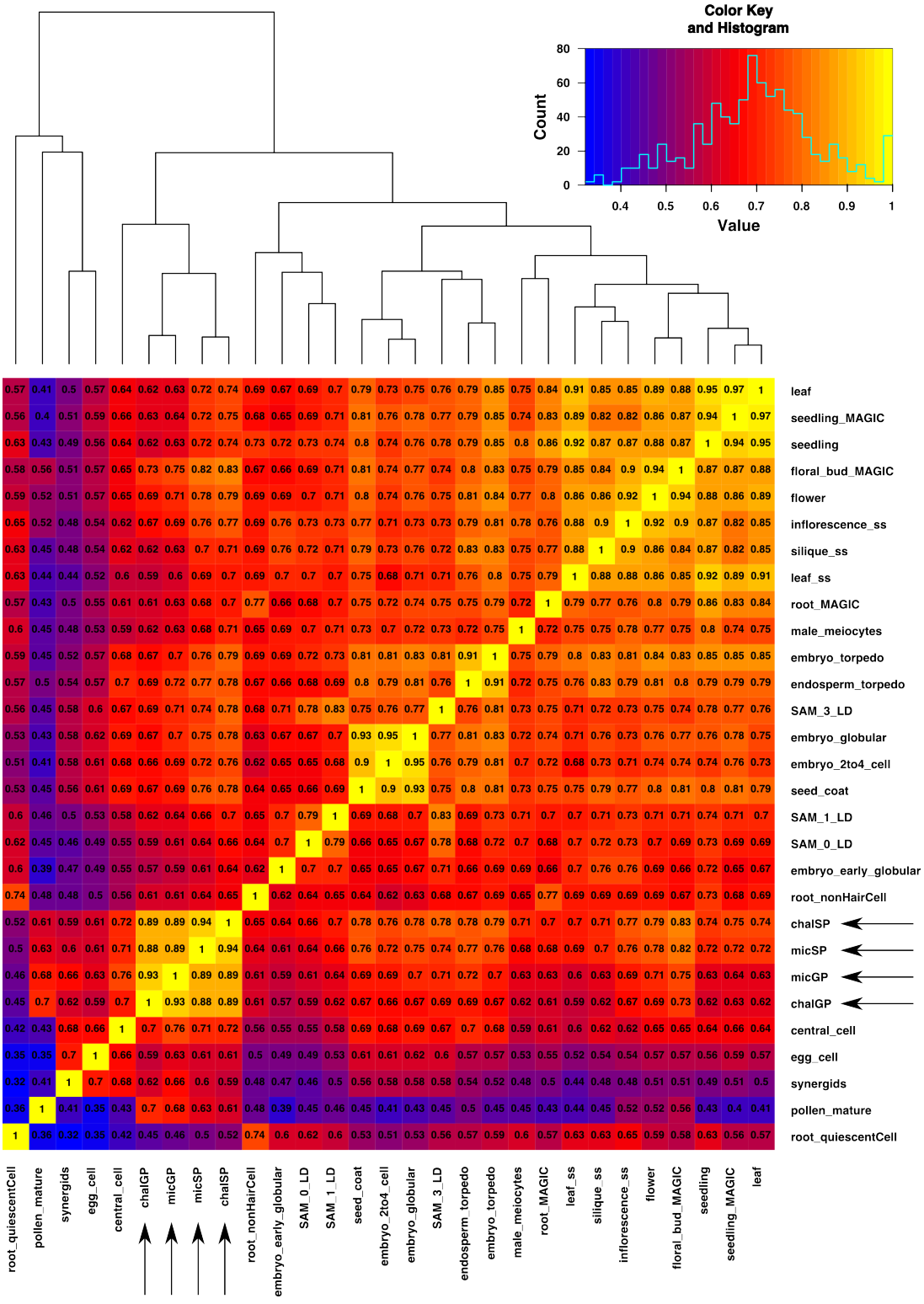


Figure S2: Heatmap illustrating the similarity between different tissue and cell types (Pearson correlation coefficients). Tissue and cell types were clustered based on their average gene expression profile using Pearson correlation and hierarchical agglomerative clustering.

Comparison of alignment statistics

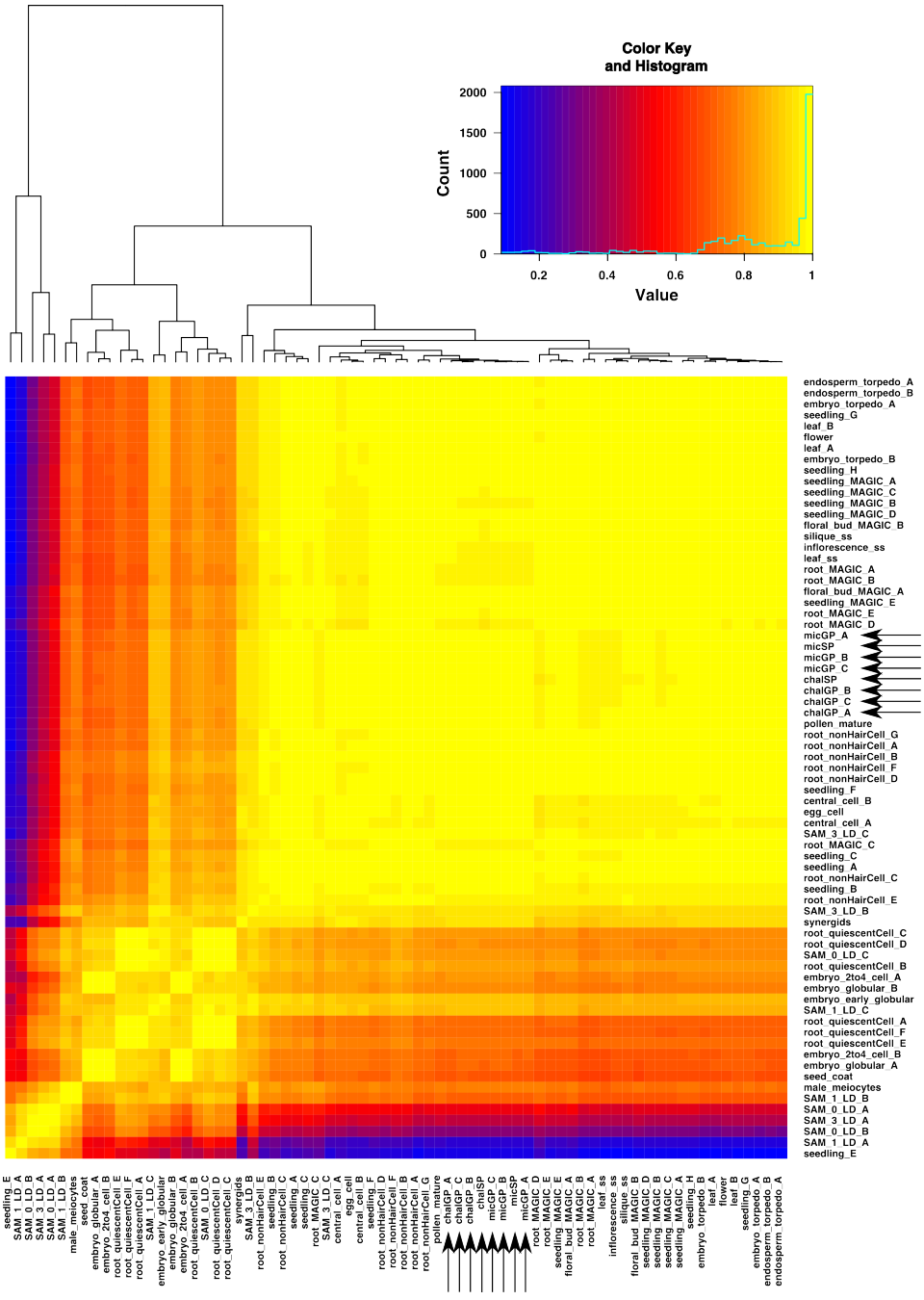


Figure S3: Heatmap illustrating the similarity between different samples in terms of their basic alignment statistics. Each alignment of a read can be characterized by several aspects. First, a read may align uniquely or multiple times to the genome. The alignment can map either to an intergenic region or to a gene. In both cases, the alignment may be gapped. Hits mapping to genes can be separated into intronic (i.e. not matching to known exons/splice junctions), exonic, or covering a splice junction. The latter two can in addition be separated into ambiguous (mapping to more than one gene) or unambiguous (mapping to only one gene). Samples were clustered using Pearson correlation and hierarchical agglomerative clustering. The dissimilarity between some of the samples was due to differences in the number of reads with multiple alignments and reads originating from intergenic regions (data not shown).

Genes significantly enriched in the female gametophyte

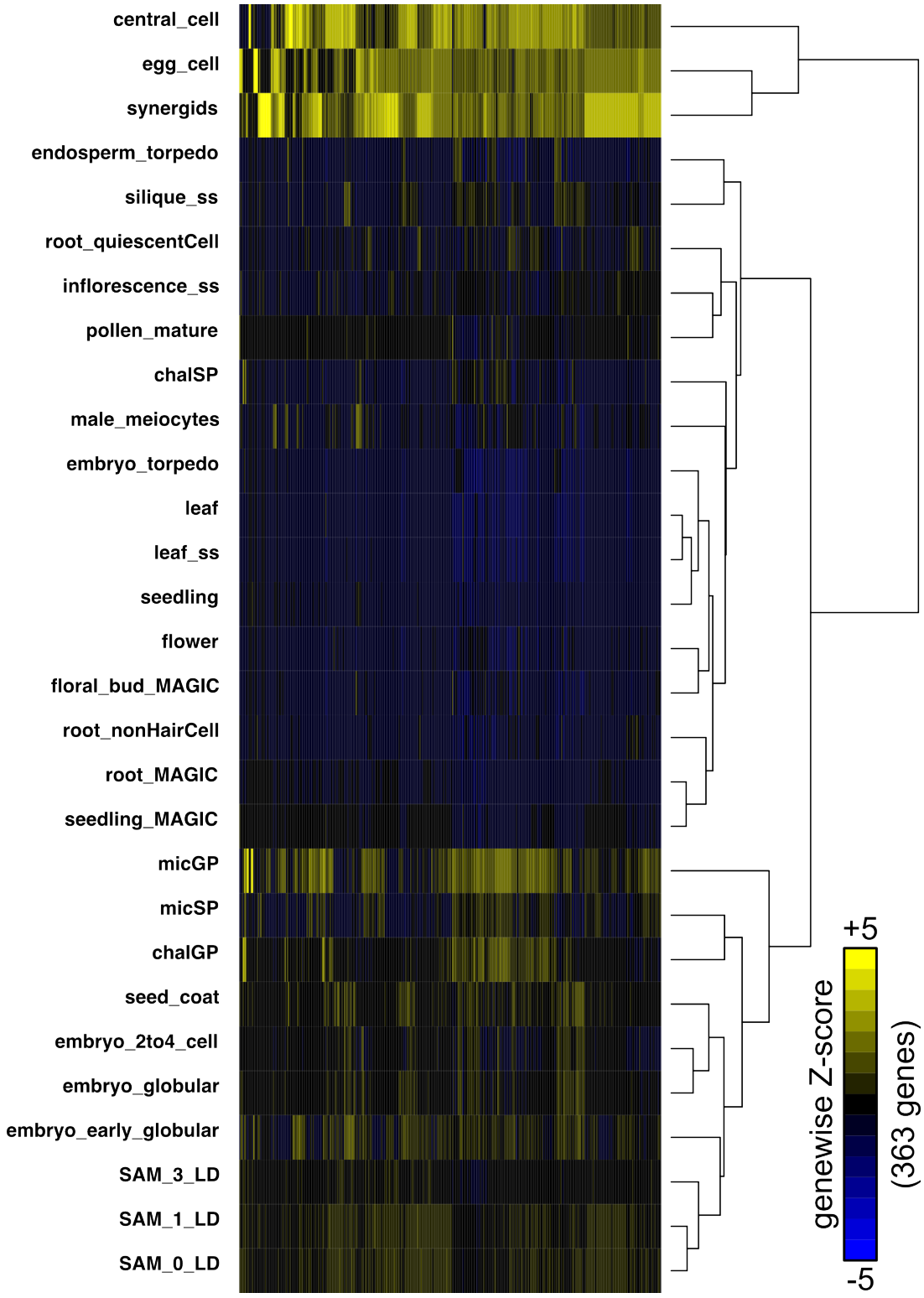


Figure S4: Expression values of genes preferentially expressed in the female gametophyte are summarized in a heatmap (blue/yellow: low/high expression values). Genewise Z-scores were calculated with the average of scaled, $\log_2$-transformed count data (mean($\log_2$(equalized counts + 1))). Cell/tissue types and genes were clustered using euclidean distance and hierarchical agglomerative clustering.

Expression pattern of small nucleolar RNAs (snoRNAs)

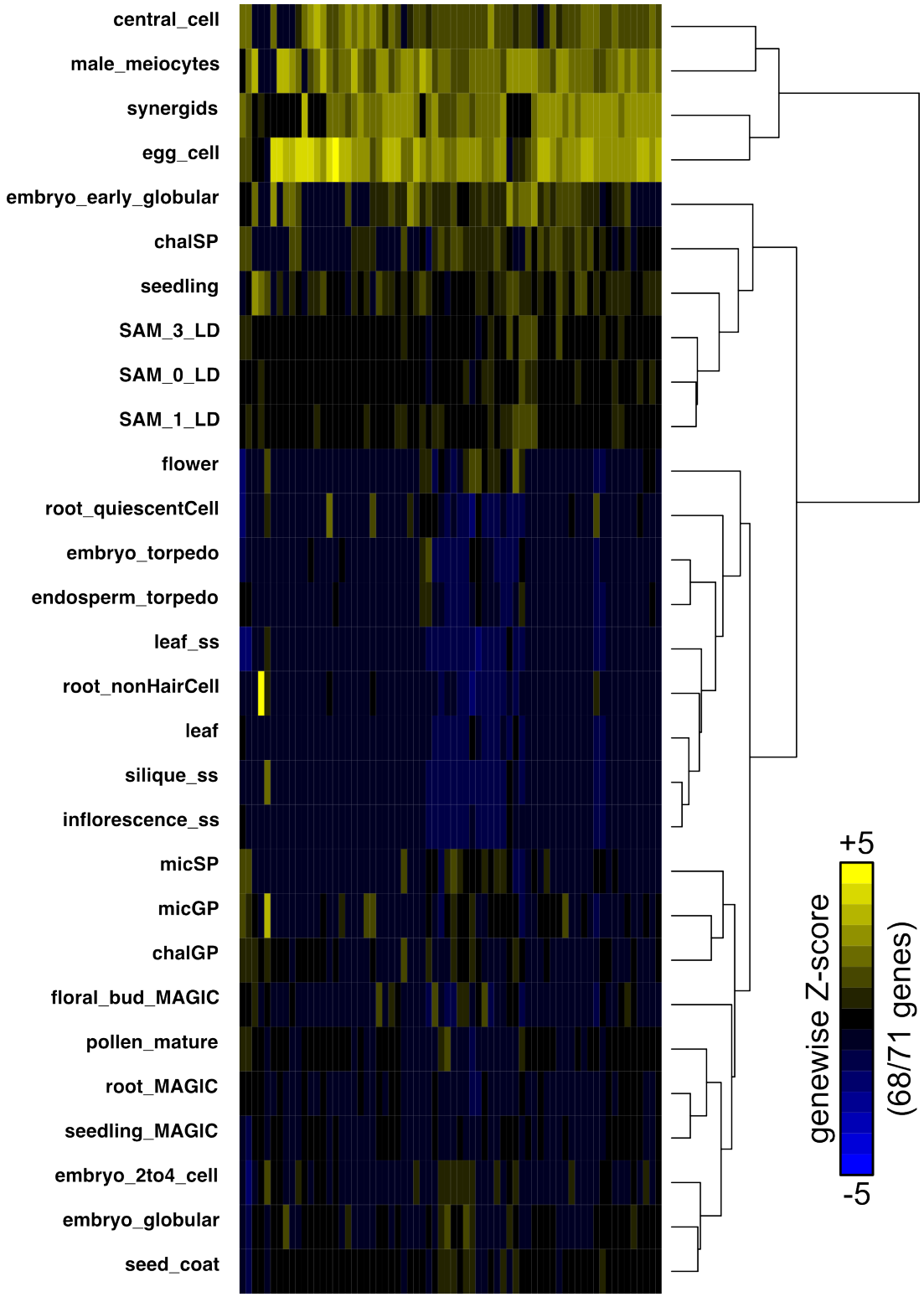


Figure S5: Expression values 68/71 snoRNA genes are summarized in a heatmap (blue/yellow: low/high expression values, three genes were not sequenced in any library). Genewise Z-scores were calculated with the average of scaled, $\log_2$-transformed count data (mean($\log_2$(equalized counts + 1))). Cell/tissue types and genes were clustered using euclidean distance and hierarchical agglomerative clustering.

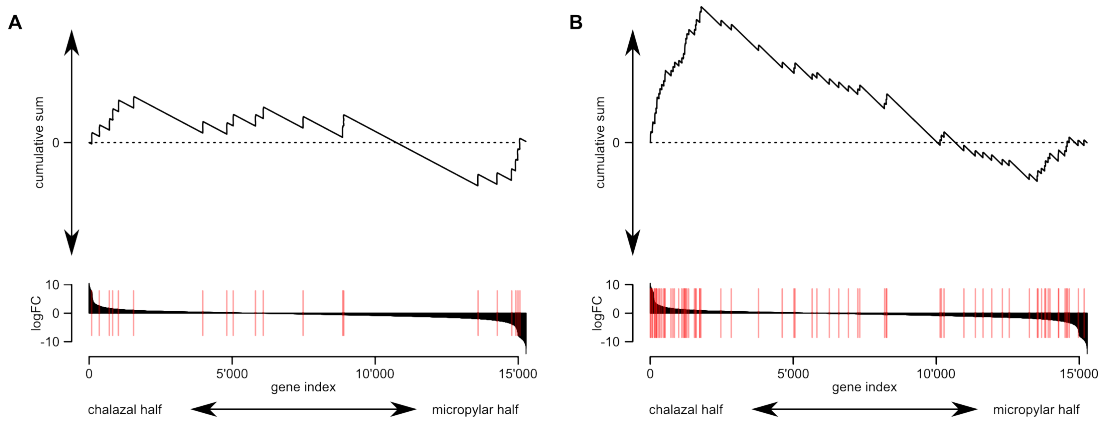


Figure S6: Test for enrichment of genes responsive to auxin (A) or cytokinin (B) in one of the two halves of the syncytial FG. Line graph on top shows the cumulative sum calculated along the list of genes sorted according to their logFC between the two halves of the syncytial FG (see Material and Methods for details). For a random set of genes, the cumulative sum would fluctuate around zero and the maximum deviation from zero (i.e. the enrichment score ES) would be small. Enrichment of the genes in either half of the FG would be visible as a cluster of genes on the top or the bottom of the sorted genes and a corresponding high ES. An empirical P-value can be calculated by comparing the observed ES to an empirical null-distribution of ES obtained through random sampling. The distribution of the 20 tested auxin-responsive genes does not differ significantly from a random distribution. In contrast, the 73 tested cytokinin-responsive genes show a highly significant enrichment in the chalazal half of the syncytial FG (P-value $<$ 0.0001). Red lines mark the genes in the gene set of interest.

RNA-Seq data

Table S1: Data used for RNA-Seq analysis. I for Illumina, S for SOLiD, SE and PE for single- and paired-end, respectively. 4nES: four-nucleate embryo sac, SAM: shoot apical meristem, SD/LD: short/long day conditions. All PE samples were treated as SE (reverse reads were removed).

accession	sample	type	rlen (bp)	comments
SRX681835	micGP_A	S_SE	50	micropylar half of the 4nES
SRX681839	micGP_B	S_PE	50/35	micropylar half of the 4nES
SRX681840	micGP_C	S_PE	50/35	micropylar half of the 4nES
SRX681838	chalGP_A	S_SE	50	chalazal half of the 4nES
SRX681841	chalGP_B	S_PE	50/35	chalazal half of the 4nES
SRX681842	chalGP_C	S_PE	50/35	chalazal half of the 4nES
SRX681833	micSP	S_SE	50	sporophyte around micGP
SRX681834	chalSP	S_SE	50	sporophyte around chalGP
SRX063784 [1]	male_meiocytes	I_SE	36	
SRX107320 [2]	embryo_2to4_cell_A	S_SE	50	cross: Col x Ler
SRX037074 [2]	embryo_2to4_cell_B	S_SE	50	cross: Ler x Col
SRX037075 [2]	embryo_globular_A	S_SE	50	cross: Ler x Col
SRX107220 [2]	embryo_globular_B	S_SE	50	cross: Ler x Col
SRX155046 [2]	seed_coat	S_SE	50	embryo at 2-4 cell stage
SRX032219 [3]	embryo_early_globular	I_SE	36	
SRX082186 [4]	embryo_torpedo_A	I_SE	36&48	cross: Col x Ler
SRX082187 [4]	endosperm_torpedo_A	I_SE	36&48	cross: Col x Ler
SRX082188 [4]	embryo_torpedo_B	I_SE	36&48	cross: Ler x Col
SRX082189 [4]	endosperm_torpedo_B	I_SE	36&48	cross: Ler x Col
SRX111836 [5]	SAM_0_LD_A	I_SE	36	from SD to LD for 0 d
SRX111837 [5]	SAM_0_LD_B	I_SE	36	from SD to LD for 0 d
SRX111838 [5]	SAM_0_LD_C	I_PE	36/36	from SD to LD for 0 d
SRX111839 [5]	SAM_1_LD_A	I_SE	36	from SD to LD for 1 d
SRX111840 [5]	SAM_1_LD_B	I_SE	36	from SD to LD for 1 d
SRX111841 [5]	SAM_1_LD_C	I_PE	36/36	from SD to LD for 1 d
SRX111842 [5]	SAM_3_LD_A	I_SE	36	from SD to LD for 3 d
SRX111843 [5]	SAM_3_LD_B	I_SE	36	from SD to LD for 3 d
SRX111844 [5]	SAM_3_LD_C	I_PE	36/36	from SD to LD for 3 d
SRX275909 [6]	pollen_mature	I_SE	75	Col-0
SRX076053 [7]	central_cell_A	S_SE	50	Ler-0
SRX076054 [7]	central_cell_B	S_SE	50	Ler-0
SRX376914 [8]	egg_cell	S_SE	50	Ler-0
SRX376915 [8]	synergids	S_SE	50	Ler-0
SRX389798 [9]	root_MAGIC_A	I_SE	100	Col-0
SRX389799 [9]	root_MAGIC_B	I_SE	100	Col-0
SRX389808 [9]	root_MAGIC_C	I_SE	100	Ler-0
SRX389809 [9]	root_MAGIC_D	I_SE	100	Ler-0
SRX084369 [9]	root_MAGIC_E	I_SE	80	Col-0
SRX084370 [9]	floral_bud_MAGIC_A	I_SE	80	Col-0
SRX389757 [9]	floral_bud_MAGIC_B	I_SE	80	Ler-0
SRX083180 [9]	seedling_MAGIC_A	I_SE	80	Col-0
SRX083181 [9]	seedling_MAGIC_B	I_SE	80	Col-0
SRX083190 [9]	seedling_MAGIC_C	I_SE	80	Ler-0
SRX083191 [9]	seedling_MAGIC_D	I_SE	80	Ler-0
SRX084368 [9]	seedling_MAGIC_E	I_SE	80	Col-0
ERR229852 [10]	root_nonHairCell_A	I_SE	44	isolated with pGL2
ERR229830 [10]	root_nonHairCell_B	I_SE	44	isolated with pGL2
ERR229844 [10]	root_nonHairCell_C	I_SE	44	isolated with pGL2
ERR229827 [10]	root_nonHairCell_D	I_SE	52	isolated with pGL2
ERR229856 [10]	root_nonHairCell_E	I_SE	52	isolated with pGL2
ERR229850 [10]	root_nonHairCell_F	I_SE	52	isolated with pGL2
ERR229841 [10]	root_nonHairCell_G	I_SE	52	isolated with pGL2
ERR229848 [10]	root_quiescentCell_A	I_SE	52	isolated with pWOX5

ERR229849 [10]	root_quiescentCell_B	I_SE	52	isolated with pWOX5
ERR229845 [10]	root_quiescentCell_C	I_SE	52	isolated with pWOX5
ERR229855 [10]	root_quiescentCell_D	I_SE	52	isolated with pWOX5
ERR229826 [10]	root_quiescentCell_E	I_SE	52	isolated with pWOX5
ERR229853 [10]	root_quiescentCell_F	I_SE	52	isolated with pWOX5
SRX336073 [11]	inflorescence_ss	I_SE	100	strand-specific
SRX336074 [11]	leaf_ss	I_SE	100	strand-specific
SRX336075 [11]	silique_ss	I_SE	100	strand-specific
SRX151565 [12]	seedling_A	I_SE	55	Col-0
SRX151566 [12]	seedling_B	I_SE	48	Col-0
SRX151567 [12]	seedling_C	I_SE	48	Col-0
SRX123429 [13]	leaf_A	I_SE	50	Col-0
SRX123433 [13]	leaf_B	I_SE	50	Col-0
SRX140484 [13]	flower	I_SE	42	Col-0
SRX150068 [14]	seedling_E	I_SE	100	Col-0
SRX150069 [14]	seedling_F	I_SE	50	Col-0
SRX150070 [14]	seedling_G	I_SE	50	Col-0
SRX150071 [14]	seedling_H	I_SE	50	Col-0

Expression data of all samples processed in this study

Table S2: The table (zip-file) contains a tab-separated table with the number of hits per gene for all samples listed in Table S1.

Alignment statistics for all RNA-Seq samples

Table S3: The table (txt-file) contains basic alignment statistics for all RNA-Seq samples used in this study. The data were visualized in Figure S3.

GO terms specifically enriched in the developing female gametophyte

Table S4: The table (txt-file) contains all genes expressed within the developing female gametophyte, which belong to one of the GO-terms found to be specifically enriched only in the transcriptome of the developing female gametophyte (GO:0051302, GO:0009567, and GO:0033013).

Genes inolved in cell fate determination and polarity of the embryo sac

Table S5: Genes playing a role in polarity, determination of cell fate, and gametophyte development in general were inferred from mutants described in the literature and tested for their expression pattern in the developing female gametophyte. Genes likely important at very early stages (e.g., meiotic genes) or for a specific function of a mature cell (e.g., genes involved in pollen tube attraction) were ignored. Candidates (EMBRYO SAC DEVELOPMENTAL ARREST) from [15] were only taken if they were characterized in greater detail (MATERNAL EFFECT EMBRYO ARREST and UNFERTILIZED EMBRYO SAC were not included as they are likely to act relatively late during FG development). Note that the citation numbering is different from the one in the main text.

gene	AGI	detected	enriched in	reference
AGL23	*AT1G65360*	n	n	[16]
AGL61/DIA	*AT2G24840*	y	n	[17, 18]
AGL62	*AT5G60440*	n	n	[19]
AGL80	*AT5G48670*	y	n	[20]
AGP18	*AT4G37450*	y	n	[21]
AHK2	*AT3G29350*	y	n	[22]
AHK3	*AT5G39340*	y	n	[22]
AHK4	*AT3G16360*	n	n	[22]
AMP1	*AT3G54720*	y	mic	[23]
APC2	*AT2G04660*	y	n	[24]
ATO	*AT5G06160*	y	n	[25]
CHR11	*AT3G06400*	y	n	[26]
CKI1	*AT2G47430*	y	n	[27, 28]
CLO/GFA1	*AT1G06220*	y	n	[29]
DME	*AT5G04560*	y	n	[30]
EDA15	*AT4G14790*	y	n	[15]
FIS2	*AT2G35670*	y	n	[31]
FU	*AT1G50240*	y	n	[32]
GEX3	*AT5G16020*	y	n	[33]
GFA2	*AT5G48030*	y	n	[34]
GPT1	*AT5G54800*	y	n	[35]
HDA7	*AT5G35600*	n	n	[36]
LIS	*AT2G41500*	y	n	[37]
MAA3	*AT4G15570*	y	mic	[38]
MSI1	*AT5G58230*	y	n	[39]
MYB64	*AT5G11050*	y	mic	[40]
MYB98	*AT4G18770*	y	mic	[41]
MYB119	*AT5G58850*	y	mic	[40]
NACK1	*AT1G18370*	y	n	[42]
NACK2	*AT3G43210*	y	n	[42]
NOMEGA	*AT1G78770*	y	n	[43]
OFP5	*AT4G18830*	n	n	[44]
RBR1	*AT3G12280*	y	n	[45]
RPT5a	*AT3G05530*	y	n	[46]
RPT5b	*AT1G09100*	y	n	[46]
SWA1	*AT2G47990*	y	n	[47]
SWA3	*AT1G16280*	y	n	[48, 49]
SYCO	*AT2G31170*	y	n	[50]
TUBG1	*AT3G61650*	y	n	[51]
TUBG2	*AT5G05620*	n	n	[51]
VDD	*AT5G18000*	y	n	[52]

Genes highly enriched in the developing female gametophyte

Table S6: A table (txt-file) with the 101 genes identified to highly enriched within the developing female gametophyte compared to all other samples listed in Table S1 containing gene type, short/long gene description, geneFamily, GO-term (biological process, molecular function, and cellular components) annotations, PFAM annotations, and InterPro (family, domain, conserved site, binding site, active site, repeat, post-translational modification) annotations.

GO terms significantly enriched in the female gametophyte

Table S7: Significantly enriched GO terms in the set of genes specific to the female gametophyte (including the synergids, egg, and central cell samples) compared to all genes expressed within the tested samples. Noteworthy, the terms "GO:2000008" and "GO:0080155" only comprise genes beloning to a family of genes with a domain of unknown function (DUF1278). Likewise, the term "GO:0000154" includes soley small nucleolar RNA (snoRNA) genes. GO-term enrichment was calculated with topGO [53] using the "weight" algorithm and Fisher's exact test. Obs/exp: number of genes observed/expected within this term.

term	obs	exp	*P*-value	description
GO:2000008	5	0.04	2.0E-11	regulation of protein localization to cell surface
GO:0000154	10	0.51	8.8E-11	rRNA modification
GO:0080155	5	0.06	1.1E-09	regulation of double fertilization forming a zygote and endosperm
GO:0010183	4	0.12	4.7E-06	pollen tube guidance
GO:0043086	4	0.70	0.0053	negative regulation of catalytic activity
GO:0048240	1	0.01	0.0073	sperm capacitation
GO:0051938	1	0.01	0.0073	L-glutamate import
GO:0043091	1	0.01	0.0146	L-arginine import
GO:0009405	1	0.02	0.0218	pathogenesis

Genes differentially expressed within the developing female gametophyte (polarized genes)

Table S8: A table (txt-file) with the 615 genes identified to be differentially expressed between the two cell halves containing average expression values, standard deviations, FDR-adjusted P-values, gene type, short/long gene description, geneFamily, GO-term (biological process, molecular function, and cellular components) annotations, PFAM annotations, and InterPro (family, domain, conserved site, binding site, active site, repeat, post-translational modification) annotations.

GO-term enrichment within the list of polarized genes.

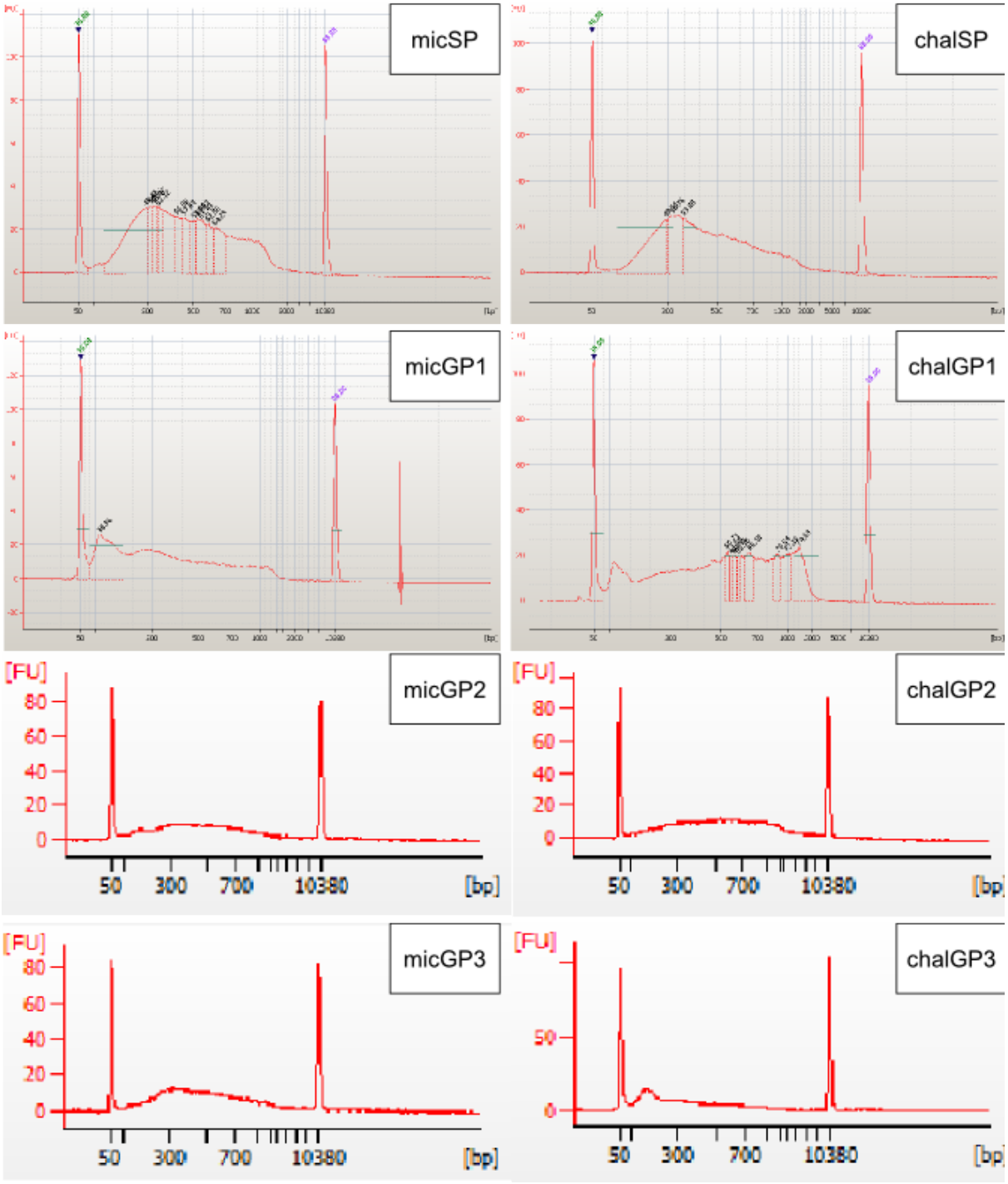
Table S9: Significantly enriched GO terms in the list of polarized genes compared to all genes having at least five reads within one of the six gametophytic samples. GO-term enrichment was calculated with topGO [53] using the "weight" algorithm and Fisher's exact test. Obs/exp: number of genes observed/expected within this term.

term	obs	exp	*P*-value	description
GO:0048868	26	7.97	1.3e-07	pollen tube development
GO:0080155	5	0.23	9.4e-07	regulation of double fertilization forming a zygote and endosperm
GO:2000008	4	0.14	3.1e-06	regulation of protein localization to cell surface
GO:0030048	11	2.16	9.3e-06	actin filament-based movement
GO:0042545	23	9.37	0.00013	cell wall modification
GO:0006436	2	0.06	0.00080	tryptophanyl-tRNA aminoacylation
GO:0009958	4	0.54	0.00177	positive gravitropism
GO:0043086	8	2.53	0.00250	negative regulation of catalytic activity
GO:0035725	5	1.05	0.00372	sodium ion transmembrane transport
GO:0006751	2	0.11	0.00464	glutathione catabolic process
GO:0009082	3	0.43	0.00801	branched-chain amino acid biosynthetic process
GO:0006949	3	0.57	0.01808	syncytium formation
GO:0016571	15	8.26	0.01950	histone methylation
GO:0071732	2	0.23	0.02009	cellular response to nitric oxide
GO:0006828	2	0.23	0.02009	manganese ion transport
GO:0015691	2	0.23	0.02009	cadmium ion transport
GO:0046274	2	0.26	0.02535	lignin catabolic process
GO:0034775	1	0.03	0.02838	glutathione transmembrane transport
GO:0033473	1	0.03	0.02838	indoleacetic acid conjugate metabolic process
GO:0048240	1	0.03	0.02838	sperm capacitation
GO:0030245	1	0.03	0.02838	cellulose catabolic process
GO:0006907	1	0.03	0.02838	pinocytosis
GO:0045697	1	0.03	0.02838	regulation of synergid differentiation
GO:0015678	1	0.03	0.02838	high-affinity copper ion transport
GO:0043693	1	0.03	0.02838	monoterpene biosynthetic process
GO:0007267	6	2.36	0.03055	cell-cell signaling
GO:0009641	2	0.34	0.04395	shade avoidance

S1 File: Library quality controls

The file contains the results from the cDNA library control experiments (size distribution of fragments and approximate concentration of selected genes).

Distribution of the size of cDNA libraries



Size distributions of cDNA libraries were measured using Agilents Bioanalyzer. Sizes are given on the x-axis, arbitrary fluorescence units corresponding to the amount of fragments with the respective size on the y-axis. Peaks at the left and right end correspond to the size markers. SP/GP for sporophyte/gametophyte and mic/chal for micropylar and chalazal halves.

Quality control of the cDNA libraries (yield)

Relative concentration of cDNA from selected genes in the cDNA libraries (water as negative control) was measured using qPCR. Given are the average Ct values from two technical replications. The value corresponds to the cycle at which the amplification product of the respective cDNA could be detected (lower value means higher relative concentration). In case no amplification product could be detected after 45 cycles, the value was set to "NA". The asterik marks the case where the amplification product could only be detected in one of the technical replicates. SP/GP for sporophyte/gametophyte and mic/chal for micropylar and chalazal halves.

sample	*ACT2*	*ACT11*	$EF-1\alpha$	*UBC9*
water	NA	NA	NA	38*
micSP	25.0	22.0	24.0	23.5
micGP1	29.5	25.0	22.0	21.0
micGP2	26.0	24.0	25.0	23.0
micGP3	26.0	24.5	23.0	24.0
water	NA	NA	NA	39*
chalSP	25.0	21.5	26.5	25.0
chalGP1	28.5	23.0	23.0	21.5
chalGP2	25.0	22.5	22.0	22.5
chalGP3	27.0	24.0	24.0	24.5

S2 File: Primer sequences

The file contains all primer sequences used in this study.

Primers for *in situ* hybridization

Primers used for cloning the probes for RNA *in situ* hybridization.

AGI	forward/reverse sequence	bp
AT2G38750	CCCCACAAGCAACTTGCGCTGA	243
	TCCCGCTTCAATACTGCGGTGG	
AT4G35165	CCACCGCCCCGCTCACAATT	255
	AGAATTGGGGACGATGTGCGC	

Primers for promoter-reporter constructs

Primers used for cloning the promoter fragments for the promoter-reporter constructs. Fragments were cloned into the target vector using ligation-independent cloning (forward adapter: TAGTTGGAATGGGTTC, reverse adapter: TTATGGAGTTGGGTTCGAA).

AGI	forward/reverse sequence	bp
AT5G52975	GAATAAGATATTTCAAATTGTAAC	287
	CTTTCTTTGTTAACCCTTTGAT	
AT1G52970	GAAGAAACAGAGTGGCTTCG	348
	TTTCTTTTTCTTGTAATGAAGAAG	
AT5G54062	TTCTCTTTTGTTAATTTCTAAGTT	487
	AGTGACTCTAGTGATCTTTTC	
AT1G74480	ACTTCATTAATAACTTATGATTAAT	521
	TGTAAAGAAAGTGAGAGAGATA	
AT2G28400	TTAATTTGATTTCTCATTTTGAATG	1378
	TTCAAGAAGAAGAAGAAATGAAA	
AT2G14378	GAAAATAAAAACCTTTACTATGAT	340
	CTCTCTTTGACTTGTATTTGC	
AT5G39420	GAGTGAAATTGTAATTTAAGGAA	1750
	TAGGCCAAAAGATTAAATGGG	
AT1G70540	CTGGGTTTAGTGGTTAAGCT	2000
	TTTGGGGTTATTGATGCTAAC	

Primers for ectopic expression constructs

Primers used for cloning the promoter and terminator regions of *At4g05440* to construct pMWS14 driving embryo sac specific expression of the gateway-casette. Region-specific sequences are in uppercase, *Asc*I and *Pac*I resctriction sites (+2 bps) in lowercase, respectively.

region	forward/reverse sequence	bp
promoter	aaggcgcgccGTGGCACATTCTAATTGGGTAAG	1'982
	aaggcgcgccCGCGATTAACGAATTCGTTGTAG	
terminator	ccttaattaaCAGAGTTCTGATGAAGTTGCTTGA	590
	ccttaattaaGAATTAAACGCAGTTTATCATAGAGAA	

Primers used for cloning the truncated genomic sequences for the ectopic expression studies. Primers were cloned into the target vector using Gateway® cloning (forward(*attB1*): AAAAAGCAGGCTTC, reverse(*attB2*): AGAAAGCTGGGTT). Genes with an asterisk were only identified as differentially expressed with an older version of edgeR using tagwise dispersion estimates but not with the newer version used in the main text.

constructID	AGI	forward/reverse sequence	bp
ox1	*AT2G39851*	ATGACTTCCCAAATATTGTTGA	831
		TTAATATCCTTTGCCATGGCT	
ox2	*AT3G09160*	ATGGACGAAATCGCCAACAA	651
		TCAACTCGCACTTGATATCAG	
ox4*	*AT4G14660*	ATGTTTCTCAAAGTCCAATTAC	537
		TCACTCTTCAGATAATGGTCC	
ox5	*AT1G70540*	ATGTCTACAAATCTCCACCTT	504
		TCATTGTTTGACAAGGGTAGA	
ox6*	*AT2G34440*	ATGGGTCGGAGAAAGATCAA	519
		CTATTTTCTGGAGAGAAGCAT	
ox7	*AT2G28400*	ATGGCGACGAGCAAGTGC	489
		TTAATCTTCGATCCCTCTAATT	
ox8	*AT5G15960*	ATGTCAGAGACCAACAAGAAT	482
		CTACTTGTTCAGGCCGGTC	
ox9	*AT5G60880*	ATGGCTTCACAGTGGACAAT	1151
		TCAGAATCTACAACATTGGAAC	
ox10	*AT5G11050*	ATGGAAGAACAGAAAATTCAAG	1492
		CTAGAACAATGGGAACCAAAT	
ox11	*AT5G58850*	ATGGAAGACCGACGCCTTG	1497
		CTAGACCAACGGGAAGCAG	
ox12	*AT5G41070*	ATGTATAAGAATCAGCTTCAAG	1370
		CTAACTATCATGGGTTTGATC	
ox13	*AT1G74480*	ATGGCTGATCACACAACCAA	1147
		TCACAAACCACTAGTAAATTCA	
ox14	*AT2G02240*	ATGTGTGGACAACATTACACG	1244
		TTACTGGCTTTTCGTAGGGC	
ox15	*AT4G23895*	ATGGATGATGGTACTCCTAAG	1264
		TTATTCTGGAGCTTCGACTG	
ox16	*AT3G57040*	ATGGGTATGGCAGCAGAATC	1170
		TCAGACAGCGGTTGCGATA	
ox17	*AT3G56850*	ATGGATTCTCAGAGGGGTAT	1341
		TCAGAAAGGAGCCGAGCTT	
ox18	*AT1G76370*	ATGAGGTGTTTCTCTTGTCTC	1580
		TTAATAACTCTGTTTTGTTTCCC	
ox19	*AT3G10560*	ATGTCATTCTTATCATTTTCTCC	1545
		TCATTGTACCCTAGGTCTCA	
ox20	*AT5G05070*	ATGCAGAGGGAGAGAATGAG	1546
		TTACTTGGGCAAGTCTAGTTG	

ox21	*AT3G56920*	ATGTCTTCTCAGAATCTTGAAC TTACCGTTCTCTAGCTTCGG	1608
ox22	*AT1G27370*	ATGGACTGCAACATGGTATC TCAGATGAAATGACTAGGGAA	1544
ox23	*AT2G44910*	ATGGGGGAAAGAGATGATGG CTAGCGACCTGATTTTTGCT	1639
ox24	*AT4G22770*	ATGGAGACTACCGGAGAAG TCACGTCAAAGTGATATTAAAG	1561
ox25*	*AT2G05160*	ATGAATTTCACAGAATCAATGAA CTATGTAACCGTTGAAATCTC	1921
ox26	*AT4G31610*	ATGGCGGATCCACCACATT TCAAACCAGATTACTGCTGAG	2077
ox27	*AT5G39420*	ATGGGTTGCATCAGCTCCA TTATCTGTCATCTTGTTCTGC	2698
ox29	*AT5G43630*	ATGGGAGATGGAGATGAGC CTAAAAGCCTAACATTTTTCTC	2947
ox30*	*AT1G61370*	ATGGGAAAGATTGGTATTGTTT TTATCGACCAACTATAGCAGT	3052
ox32	*AT1G68570*	ATGGAGGAGCAAAGCAAGAA TCATTCATCAACTAAACTCCTA	3300
ox33*	*AT1G22760*	ATGGCGGCGGCGGTTGC TCAGTCGGTGGTAGAGAAAC	3201
ox34	*AT2G37010*	ATGAGAGTTAGGGTTGATGTT TTATTTCTTCTGGAATGTAACC	4440
ox35	*AT5G48600*	ATGGAGGAAGATGAGCCAAT CTAAGCAGGAGTTTTCTGAC	7184
ox37	*AT4G38070*	ATGGAGAAGGTTTATGAAGAG TCATTGCTGCTGAATTAACAAC	6035

References

[1] Chen C, Farmer AD, Langley RJ, Mudge J, Crow JA, et al. (2010) Meiosis-specific gene discovery in plants: RNA-Seq applied to isolated *Arabidopsis* male meiocytes. BMC Plant Biology 10: 280.

[2] Autran D, Baroux C, Raissig MT, Lenormand T, Wittig M, et al. (2011) Maternal epigenetic pathways control parental contributions to *Arabidopsis* early embryogenesis. Cell 145: 707-719.

[3] Nodine MD, Bartel DP (2010) MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. Genes & Development 24: 2678-2692.

[4] Gehring M, Missirian S V Henikoff (2011) Genomic analysis of parent-of-origin allelic expression in *Arabidopsis thaliana* seeds. PLOS ONE 6: e23687.

[5] Torti S, Fornara F, Vincent C, Andrés F, Nordström K, et al. (2012) Analysis of the *Arabidopsis* shoot meristem transcriptome during floral transition identifies distinct regulatory patterns and a leucine-rich repeat protein that promotes flowering. The Plant Cell 24: 444-462.

[6] Loraine AE, McCormick S, Estrada A, Patel K, Qin P (2013) RNA-Seq of *Arabidopsis* pollen uncovers novel transcription and alternative splicing. Plant Physiology 162: 1092-1109.

[7] Schmid MW, Schmidt A, Klostermeier UC, Barann M, Rosenstiel P, et al. (2012) A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing. PLOS ONE 7: e29685.

[8] Schmidt A, Schmid MW, Klostermeier UC, Qi W, Guthörl D, et al. (2014) Apomictic and sexual germline development differ with respect to cell cycle, transcriptional, hormonal and epigenetic regulation. PLOS Genetics 10: e1004476.

[9] Gan X, Stegle O, Behr J, Steffen JG, Drewe P, et al. (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477: 419-423.

[10] Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, et al. (2013) Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods 10: 1093-1095.

[11] Wang H, Chung PJ, Liu J, Jang IC, Kean MJ, et al. (2014) Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in *Arabidopsis*. Genome Research 24: 3.

[12] Zhu Y, Rowley MJ, Böhmdorfer G, Wierzbicki AT (2013) A SWI/SNF chromatin-remodeling complex acts in noncoding RNA-mediated transcriptional silencing. Molecular Cell 49: 298-309.

[13] Ausin I, Greenberg MV, Simanshu DK, Hale CJ, Vashisht AA, et al. (2012) INVOLVED IN DE NOVO 2-containing complex involved in RNA-directed DNA methylation in *Arabidopsis*. PNAS 109: 8374-8381.

[14] Stroud H, Hale CJ, Feng S, Caro E, Jacob Y, et al. (2012) DNA methyltransferases are required to induce heterochromatic re-replication in *Arabidopsis*. PLOS Genetics 8: e1002808.

[15] Pagnussat GC, Yu HJ, Ngo QA, Rajani S, Mayalagu S, et al. (2005) Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. Development 132: 603-614.

[16] Colombo M, Masiero S, Vanzulli S, Lardelli P, Kater MM, et al. (2008) *AGL23*, a type I MADS-box gene that controls female gametophyte and embryo development in *Arabidopsis*. The Plant Journal 54: 1037-1048.

[17] Bemer M, Wolters-Arts M, Grossniklaus U, Angenent GC (2008) The MADS domain protein DIANA acts together with AGAMOUS-LIKE80 to specify the central cell in *Arabidopsis* ovules. The Plant Cell 20: 2088-2101.

[18] Steffen JG, Kang IH, Portereiko MF, Lloyd A, Drews GN (2008) AGL61 interacts with AGL80 and is required for central cell development in *Arabidopsis*. Plant Physiology 148: 259-268.

[19] de Folter S, Immink RG, Kieffer M, Parenicová L, Henz SR, et al. (2005) Comprehensive interaction map of the *Arabidopsis* MADS box transcription factors. The Plant Cell 17: 1424-1433.

[20] Portereiko MF, Lloyd A, Steffen JG, Punwani JA, Otsuga D, et al. (2006) AGL80 is required for central cell and endosperm development in *Arabidopsis*. The Plant Cell 18: 1862-1872.

[21] Acosta-Garcia G, Vielle-Calzada JP (2004) A classical arabinogalactan protein is essential for the initiation of female gametogenesis in *Arabidopsis*. The Plant Cell 16: 2614-2628.

[22] Cheng CY, Mathews DE, Schaller GE, Kieber JJ (2013) Cytokinin-dependent specification of the functional megaspore in the *Arabidopsis* female gametophyte. The Plant Journal 73: 929-940.

[23] Kong J, Lau S, Gerd J (2015) Twin plants from supernumerary egg cells in *Arabidopsis*. Current Biology 25: 225-230.

[24] Capron A, Serralbo O, Fülöp K, Frugier F, Parmentier Y, et al. (2003) The *Arabidopsis* anaphase-promoting complex or cyclosome: molecular and genetic characterization of the APC2 subunit. The Plant Cell 15: 2370-2382.

[25] Moll C, von Lyncker L, Zimmermann S, Kägi C, Baumann N, et al. (2008) *CLO/GFA1* and *ATO* are nocel regulators of gametic cell fate in plants. The Plant Journal 56: 913-921.

[26] Huanca-Mamani W, Garcia-Aguilar M, León-Martínez G, Grossniklaus U, Vielle-Calzada JP (2005) CHR11, a chromatin-remodeling factor essential for nuclear proliferation during female gametogenesis in *Arabidopsis thaliana*. PNAS 102: 17231-17236.

[27] Pischke MS, Jones LG, Otsuga D, Fernandez DE, Drews GN, et al. (2002) An *Arabidopsis* histidine kinase is essential for megagametogenesis. PNAS 99: 15800-15805.

[28] Hejátko J, Pernisová M, Eneva T, Palme K, Brzobohatỳ B (2003) The putative sensor histidine kinase CKI1 is involved in female gametophyte development in *Arabidopsis*. Molecular Genetics and Genomics 269: 443-453.

[29] Moll C, Nielsen N, Gross-Hardt R (2008) Mutants with abberant numbers of gametic cells shed new light on old questions. Plant Biology (Stuttgart, Germany) 10: 529-533.

[30] Choi Y, Gehring M, Johnson L, Hannon M, Harada JJ, et al. (2002) DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in *Arabidopsis*. Cell 110: 33-42.

[31] Chaudhury AM, Ming L, Miller C, Craig S, Dennis ES, et al. (1997) Fertilization-independent seed development in *Arabidopsis thaliana*. PNAS 94: 4223-4228.

[32] Sung AO, Johnson A, Smertenko A, Rahman D, Soon KP, et al. (2005) A divergent cellular role for the FUSED kinase family in the plant-specific cytokinetic phragmoplast. Current Biology 15: 2107-2111.

[33] Alandete-Saez M, Ron M, McCormick S (2008) *GEX3*, expressed in the male gametophyte and in the egg cell of *Arabidopsis thaliana*, is essential for micropylar pollen tube guidance and plays a role during early embryogenesis. Molecular Plant 1: 586-598.

[34] Christensen CA, Gorsich SW, Brown RH, Jones LG, Brown J, et al. (2002) Mitochondrial GFA2 is required for synergid cell death in *Arabidopsis*. The Plant Cell 14: 2215-2232.

[35] Niewiadomski P, Knappe S, Geimer S, Fischer K, Schulz B, et al. (2005) The *Arabidopsis* plastidic glucose 6-phosphate/phosphate translocator GPT1 is essential for pollen maturation and embryo sac development. The Plant Cell 17: 760-775.

[36] Cigliano RA, Cremona G, Paparo R, Termolino P, Perrella G, et al. (2013) Histone deacetylase AtHDA7 is required for female gametophyte and embryo development in *Arabidopsis*. Plant Physiology 163: 431-440.

[37] Gross-Hardt R, Kägi C, Baumann N, Moore JM, Baskar R, et al. (2007) *LACHESIS* restricts gametic cell fate in the female gametophyte of *Arabidopsis*. PLOS Biology 5: e47.

[38] Shimizu KK, Okada K (2000) Attractive and repulsive interactions between female and male gametophytes in *Arabidopsis* pollen tube guidance. Development 127: 4511-4518.

[39] Guitton AE, Berger F (2005) Loss of function of MULTICOPY SUPPRESSOR OF IRA 1 produces nonviable parthenogenetic embryos in *Arabidopsis*. Current Biology 15: 750-754.

[40] Rabinger DS, Drews GN (2013) *MYB64* and *MYB119* are required for cellularization and differentiation during female gametogenesis in *Arabidopsis thaliana*. PLOS Genetics 9: e1003783.

[41] Punwani JA, Rabiger DS, Drews GN (2007) MYB98 positively regulates a battery of synergid-expressed genes encoding filiform apparatus-localized proteins. The Plant Cell 19: 2557-2568.

[42] Tanaka H, Ishikawa M, Kitamura S, Takahashi Y, Soyano T, et al. (2004) The *AtNack1/HINKEL* and *STUD/TETRASPORE/AtNACK2* genes, which encode functionally redundant kinesins, are essential for cytokinesis in *Arabidopsis*. Genes to Cells 9: 1199-1211.

[43] Kwee HS, Sundaresan V (2003) The *NOMEGA* gene required for female gametophyte development encodes the putative APC6/CDC16 component of the anaphase promoting complex in *Arabidopsis*. The Plant Journal 36: 853-866.

[44] Pagnussat GC, Yu HJ, Sundaresan V (2007) Cell-fate switch of synergid to egg cell in *Arabidopsis eostre* mutant embryo sacs arises from misexpression of the bel1-like homeodomain gene *BLH1*. The Plant Cell 19: 3578-3592.

[45] Ebel C, Mariconti L, Gruissem W (2004) Plant retinoblastoma homologues control nuclear proliferation in the female gametophyte. Nature 429: 776-780.

[46] Gallois JL, Guyon-Debast A, Lécureuil A, Vezon D, Carpentier V, et al. (2009) The *Arabidopsis* proteasome RPT5 subunits are essential for gametophyte development and show accession-dependent redundancy. The Plant Cell 21: 442-459.

[47] Shi DQ, Liu J, Xiang YH, Ye D, Sundaresan V, et al. (2005) SLOW WALKER1, essential for gametogenesis in *Arabidopsis*, encodes a WD40 protein involved in 18S ribosomal RNA biogenesis. The Plant Cell 17: 2340-2354.

[48] Liu M, Shi DQ, Yuan L, Liu J, Yang WC (2010) *SLOW WALKER3*, encoding a putative DEAD-box RNA helicase, is essential for female gametogenesis in *Arabidopsis*. Journal of Integrative Plant Biology 52: 817-828.

[49] Huang CK, Huang LF, Huang JJ, Wu SJ, Yeh CH, et al. (2010) A DEAD-box protein, AtRH36, is essential for female gametophyte development and is involved in rRNA biogenesis in *Arabidopsis*. Plant Cell Physiology 51: 694-706.

[50] Kägi C, Baumann N, Nielsen N, Stierhof YD, Gross-Hardt R (2010) The gametic central cell of *Arabidopsis* determines the lifespan of adjacent accessory cells. PNAS 107: 22350-22355.

[51] Pastuglia M, Azimzadeh J, Goussot M, Camilleri C, Belcram K, et al. (2006) γ-tubulin is essential for microtubule organization and development in *Arabidopsis*. The Plant Cell 18: 1412-1425.

[52] Matias-Hernandez L, Battaglia R, Galbiati F, Rubes M, Eichenberger C, et al. (2010) *VERDANDI* is a direct target of the MADS domain ovule identity complex and affects embryo sac differentiation in *Arabidopsis*. The Plant Cell 22: 1702-1715.

[53] Alexa A, Rahnenführer J (2010) topGO: Enrichment analysis for Gene Ontology. R package version 2 18 0.

4 A Powerful Method for Transcriptional Profiling of Specific Cell Types in Eukaryotes: Laser-Assisted Microdissection and RNA Sequencing

The following manuscript is published in "PLOS ONE" (open access)[1]. Ueli Grossniklaus, Anja Schmidt, and I designed and conceived the study. I designed, generally performed, analyzed, and interpreted the results of the individual experiments. However, Ulrich C. Klostermeier performed the RNA-Seq library preparation, quality controls (data underlying Supplemental File S1), and the sequencing (supported by Philip Rosenstiel). Matthias Barann provided the initial data analysis used in my master thesis (not included in the manuscript). I further analyzed all data, wrote the manuscript, and created/assembled all tables and figures. Anja Schmidt critically read the manuscript and provided valuable feedback. Ueli Grossniklaus read and corrected the final draft.

[1]Schmid, MW, Schmidt, A, Klostermeier, UC, Barann, M, Rosenstiel, P, and Grossniklaus, U (2012) A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing. PLOS ONE 7: e29685.

OPEN ACCESS Freely available online

PLoS one

A Powerful Method for Transcriptional Profiling of Specific Cell Types in Eukaryotes: Laser-Assisted Microdissection and RNA Sequencing

Marc W. Schmid[1], Anja Schmidt[1], Ulrich C. Klostermeier[2], Matthias Barann[2], Philip Rosenstiel[2], Ueli Grossniklaus[1]*

1 Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zürich, Switzerland, **2** Institute of Clinical Molecular Biology, Christian-Albrechts University, Kiel, Germany

Abstract

The acquisition of distinct cell fates is central to the development of multicellular organisms and is largely mediated by gene expression patterns specific to individual cells and tissues. A spatially and temporally resolved analysis of gene expression facilitates the elucidation of transcriptional networks linked to cellular identity and function. We present an approach that allows cell type-specific transcriptional profiling of distinct target cells, which are rare and difficult to access, with unprecedented sensitivity and resolution. We combined laser-assisted microdissection (LAM), linear amplification starting from <1 ng of total RNA, and RNA-sequencing (RNA-Seq). As a model we used the central cell of the *Arabidopsis thaliana* female gametophyte, one of the female gametes harbored in the reproductive organs of the flower. We estimated the number of expressed genes to be more than twice the number reported previously in a study using LAM and ATH1 microarrays, and identified several classes of genes that were systematically underrepresented in the transcriptome measured with the ATH1 microarray. Among them are many genes that are likely to be important for developmental processes and specific cellular functions. In addition, we identified several intergenic regions, which are likely to be transcribed, and describe a considerable fraction of reads mapping to introns and regions flanking annotated loci, which may represent alternative transcript isoforms. Finally, we performed a *de novo* assembly of the transcriptome and show that the method is suitable for studying individual cell types of organisms lacking reference sequence information, demonstrating that this approach can be applied to most eukaryotic organisms.

Citation: Schmid MW, Schmidt A, Klostermeier UC, Barann M, Rosenstiel P, et al. (2012) A Powerful Method for Transcriptional Profiling of Specific Cell Types in Eukaryotes: Laser-Assisted Microdissection and RNA Sequencing. PLoS ONE 7(1): e29685. doi:10.1371/journal.pone.0029685

Editor: Shin-Han Shiu, Michigan State University, United States of America

Received August 22, 2011; **Accepted** December 1, 2011; **Published** January 26, 2012

Funding: This work was supported by the University of Zurich, and grants of the "Stiftung fuer wissenschaftliche Forschung" (through support by the "Baumgarten Stiftung") and the Swiss National Science Foundation to UG. The study was supported by Life Technologies by contributing, in part, sequencing reagents, which had no influence on the design of the study. PR is supported by the DFG Clusters of Excellence "Future Ocean" and "Inflammation at Interfaces" and the NGFN Network Genomics of Chronic Inflammatory Diseases. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: grossnik@botinst.uzh.ch

Introduction

The development of multicellular organisms is controlled by distinct cell fate decisions, which are largely mediated through the establishment of cell- or tissue-specific gene expression patterns. Spatially and temporally resolved information on gene expression patterns facilitate the identification of regulatory networks of gene activity that underly cell differentiation and functional specification. However, transcriptional profiling of specific cell types requires their isolation from an often heterogenic tissue and the determination of the transcriptome, preferentially with high resolution and completeness from ultra-low amounts of RNA (down to single cell resolution).

One method used successfully for the transcriptional profiling of distinct cell types is laser-assisted microdissection (LAM) in combination with DNA microarrays (examples [1,2] in human and [3–6] in plants). LAM allows the isolation of individual cells directly from the surrounding tissue based on histological identification with little cross-contamination (independently validated in [3]). Cell type-specific markers can assist the identification but are not required for LAM. During the procedures, alterations of cellular processes are unlikely because the tissue is rapidly fixed prior to embedding, sectioning, and LAM [7]. However, only limited amounts of RNA can be isolated from rare cell types and obtaining sufficient amounts for transcriptome analysis usually requires RNA amplification. Several methods for linear RNA amplification suitable for microarray analysis have been successfully established, leading to new insights into the transcriptional state of specific cell types [1–6]. Nonetheless, microarrays have several limitations: high background levels due to cross-hybridization, a lack of sensitivity at low and high expression levels, and reliance upon existing knowledge about the genome sequence [8]. The recently developed high-throughput sequencing of RNA using next-generation sequencing platforms (RNA-Seq) has the potential to overcome these limitations [8,9] and offers a variety of new possibilities such as the transcriptional profiling of organisms lacking sequence information [10], or the identification of novel loci, alternative splicing events [11], and sequence variation [12].

Given the advantages and opportunities offered by RNA-Seq, a combination of LAM and RNA-Seq promises to be a valuable tool for the transcriptional profiling of individual cell types. We expected that RNA-Seq would provide a more comprehensive view on the transcriptomes of specific cell types in means of completeness and complexity. That is, the detection of a larger number of expressed genes, the identification of transcripts from previously unannotated loci, and the description of genome-wide transcriptional patters. We therefore established the combination of LAM, linear RNA amplification, and RNA-Seq using the Life Technology SOLiD platform.

As a model system we used the highly inaccessible female gametophyte (embryo sac) of *Arabidopsis thaliana* (Figure 1). The embryo sacs develop within the ovules which are themselves located within the ovary of a flower. The development of an embryo sac starts with a functional megaspore (meiotic product) that undergoes three mitotic divisions in a syncytium. Nuclear migration and concomitant cellularization eventually leads to the formation of an eight-nucleate, seven-celled female gametophyte. At maturity, the embryo sac contains three distinct cell types: the synergids and the two female gametes: the egg and the central cell [13] that, following fertilization, give rise to the embryo and endosperm, respectively. These cells are therefore good examples of rare cell types which are difficult to access. The transcriptomes of these cell types have only recently been determined using LAM in combination with Affymetrix ATH1 microarrays [3], making them an ideal system to establish the combination of LAM and RNA-Seq and to compare the two transcriptome profiling techniques.

In this study, we isolated RNA from central cells collected by LAM, prepared sequencing libraries following a protocol designed for the transcriptome analysis of a single cell [14], and sequenced them using the Life Technology SOLiD platform. We estimate the number of expressed genes (defined by having at least five reads within one replicate) to be more than twice the number reported previously in a study using LAM and ATH1 microarrays [3], and identified several intergenic regions which are likely to be transcribed. We further describe a considerable fraction of reads mapping to introns and regions close to the borders of known loci, indicating extensive alterations during transcription. Finally, we performed a *de novo* assembly of the transcriptome and showed that the workflow could also be used to study organisms lacking a reference genome. Taken together, the results indicate superior performance of the workflow presented here over the frequently used approach that combines LAM with transcriptome microarrays. We believe that the approach established in this study can be used for the cell type-specific transcriptional profiling of most eukaryotic organisms, and thus, significantly contributes to the understanding of the molecular processes underlying the development of multicellular organisms.

Results and Discussion

RNA isolation, library preparation and sequencing

We used LAM to dissect the central cells of the mature embryo sac. After the isolation of the cells, we used larger sections from the remaining tissue to monitor the RNA integrity with Agilent's Bioanalyzer. As a measure for this, the machine provides the RNA Integrity Number (RIN) with a range of 1 to 10, where 10 stands for a perfect RNA sample. Using an optimized version of the protocol in [3] for sample preparation, we obtained a RIN of around 8 with minor variations between different samples (data not shown).

We aimed to analyze two biological replicates (termed CC1 and CC2). Per replicate we pooled sections from approximately 450 cells during RNA extraction. Due to the low amounts of total RNA obtained with this procedure (estimated 300–1'000 pg) amplification was required. Therefore, we used the protocol described in [14], which is designed to generate cDNA libraries suitable for SOLiD sequencing from RNA isolated from a single cell. In short, mRNA is reverse transcribed to cDNA with poly(T)-primers fused to anchor sequences for PCR amplification. After PCR amplification, cDNA is sheared and amplified again after the ligation of the sequencing adapters. To monitor the efficiency of

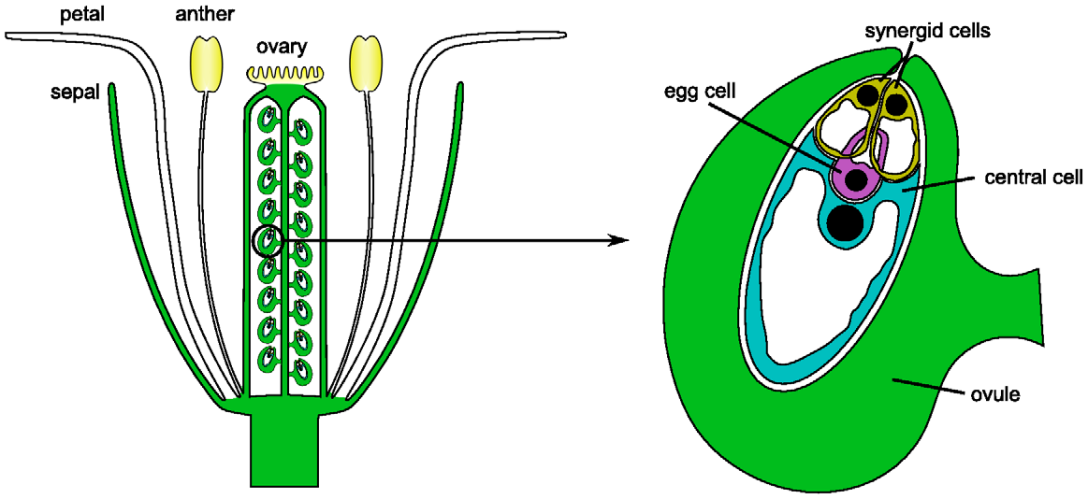


Figure 1. Schematic representation of the flower and the embryo sac of *Arabidopsis thaliana*. The flower of *Arabidopsis thaliana* consists of four whorls of organs: sepals, petals, anthers (male reproductive organs) and carpels (female reproductive organs). The carpels are fused and form the ovary, which harbors around fifty ovules. During ovule development, one embryo sac is formed within each ovule. The mature embryo sac contains three distinct cell types: the synergids and the two female gametes: the egg and the central cell [13]. The mature embryo sac of *Arabidopsis thaliana*, accession Landsberg *erecta*, is around 105 μm long and 25 μm wide [44]. The nuclei of the cells of the embryo sac are drawn as black circles, the vacuoles as white regions.
doi:10.1371/journal.pone.0029685.g001

the library preparation we measured the size of the cDNAs prior to shearing using Agilent's Bioanalyzer and the approximate concentration of cDNA from selected genes with qRT-PCR (File S1). The cDNA of the first replicate (CC1) had a slightly smaller size range and lower yield than the one from the second replicate (CC2). The size distributions of around 0.1–1.5 kb indicated shortening of the RNA fragments as compared to the average full-length transcripts from *Arabidopsis thaliana* (1.5 kb). As a consequence, the sequence coverage of longer transcripts was not uniform but shifted to the 3′ ends of the transcripts (3′ bias). The bias was likely due to the oligo-dT primed cDNA generation, which has been reported to preferentially represent the 3′ ends of transcripts when compared to direct RNA fragmentation [8,15]. However, optimized oligo-dT or direct RNA fragmentation protocols, such as described in [15], rely on mRNA enrichment and are therefore not suitable for the low amounts of total RNA obtained with the methods described here [16].

The libraries were sequenced using the SOLiD platform (version 3, Life Technology, Foster City). Each library was sequenced on one eighth of a slide resulting in a total number of 43'740'114 and 43'987'011 reads (50 bp) for the first (CC1) and the second (CC2) replicate, respectively. Potential sequencing errors were corrected using the SOLiD Accuracy Enhancement Tool (solidsoftwaretools.com/gt/project/saet). We first analyzed the data using an approach that is based on the alignment of reads to the *Arabidopsis thaliana* reference genome.

Data analysis using a reference genome

The corrected reads were aligned to the *Arabidopsis thaliana* reference genome (www.arabidopsis.org) using TopHat [17], which is designed to identify previously undescribed splice junctions. To avoid a potential underestimation of expression levels of gene family members with similar transcript sequences we allowed up to ten alignments per read. The alignments of those reads were then weighted based on the number of uniquely aligned reads in the neighborhood. By these criteria, around 42% of the reads had at least one valid alignment, corresponding to 18'907'766 (CC1) and 18'038'960 (CC2) weighted alignments (in the following sections we use "hits" as a synonym for an alignment that was weighted).

Genome-wide patterns. To get a genome-wide overview of the results, we classified the hits based on their location in the genome (Table 1). The majority of the hits was found within exons and across splice junctions (82.6%). The other hits could be divided into four groups with hits (i) mapping to intronic regions (8.5%), (ii) located at or very close (distance below 200 bp) to the borders of known loci (4.8%), (iii) overlapping with known transposable elements in the intergenic regions (1%) and (iv), isolated from any known genomic feature (3.1%). For each group we then obtained the genomic regions which were sequenced in both replicates and calculated the number of hits overlapping with these "reproducibly sequenced" regions (Table 1). Overall, the sequence coverage between the replicates was highly similar with around 97.1% of all hits overlapping with a reproducibly sequenced region. Hits specific to one replicate were likely caused by a slightly differential amplification efficiency between the replicates. Furthermore, it is also possible that a higher sequencing depth would improve the similarity between the replicates in terms of sequence coverage. Nonetheless, the high percentages clearly indicate a good reproducibility of sequence coverage on a genome-wide scale.

Overall, the percentage of non-exonic hits (in total 17.4% of all hits) is well above the percentages reported in other RNA-Seq studies on *Arabidopsis thaliana* (pool of organs and seedlings in [18]: 7%; unopened flower buds in [19]: 3.5%). An explanation for this difference may be that the annotation of the *Arabidopsis thaliana* genome is widely based on sequencing of cDNAs and expressed sequence tags (ESTs) originating from larger plant structures or whole plants. Transcripts uniquely expressed in small structures or rare cell types, such as the female gametophytic cells, were therefore less likely to be detected due to a dilution effect. Thus, the high percentage of intergenic hits in the data presented here may partly reflect transcripts or transcript isoforms specific to the central cell. Detailed analysis of transcript isoforms from known loci is, however, difficult due to the non-uniform sequence coverage. Nonetheless, we used the intergenic hits which were isolated from any known genomic feature to search for (fragments) of transcripts from previously unannotated loci. To identify these loci we used cufflinks [11], which is designed to assemble transcripts from reads which were aligned to a reference genome (with the focus on paired-end read libraries). Using single-end and unstranded reads, the program assumes uniform coverage along a transcript. It is therefore not well suited for an in-depth analysis of data generated with the methods presented here. Nonetheless, we could identify 78 (CC1) and 115 (CC2) potentially new transcripts, which were supported by one or more splice junctions (Table S1). Many of them showed a coverage pattern similar to the one observed for annotated transcripts (example in Figure 2B). These transcripts corresponded to 75 (CC1) and 112 (CC2) putative loci, in the following termed "splice-loci". To compare their genomic location between the two replicates, we calculated for each of them the overlap with a splice-locus/loci from the other replicate and counted the number of splice-loci with an overlap above a certain threshold (Figure S1). Splice-loci with very low expression values (less than 25 hits) showed a poor overlap between the two replicates, irrespective of the threshold (11% with perfect overlap

Table 1. Classification of alignments.

genomic region	CC1	CC2
genome and splice junctions (total)	18'907'766.00 (97.93%)	18'038'960.00 (96.22%)
exons and splice junctions	15'456'413.54 (98.50%)	15'069'463.75 (96.65%)
introns	1'652'728.54 (93.66%)	1'485'453.60 (92.67%)
regions flanking loci	977'004.27 (95.66%)	797'708.77 (94.22%)
transposable elements outside of loci	200'268.10 (94.34%)	166'855.87 (94.39%)
remaining intergenic regions	621'351.56 (93.23%)	519'478.01 (91.56%)

The table summarizes the number of hits found in a certain genomic region. The percentage of these hits which were overlapping with regions sequenced in both replicates are given in parentheses.
doi:10.1371/journal.pone.0029685.t001

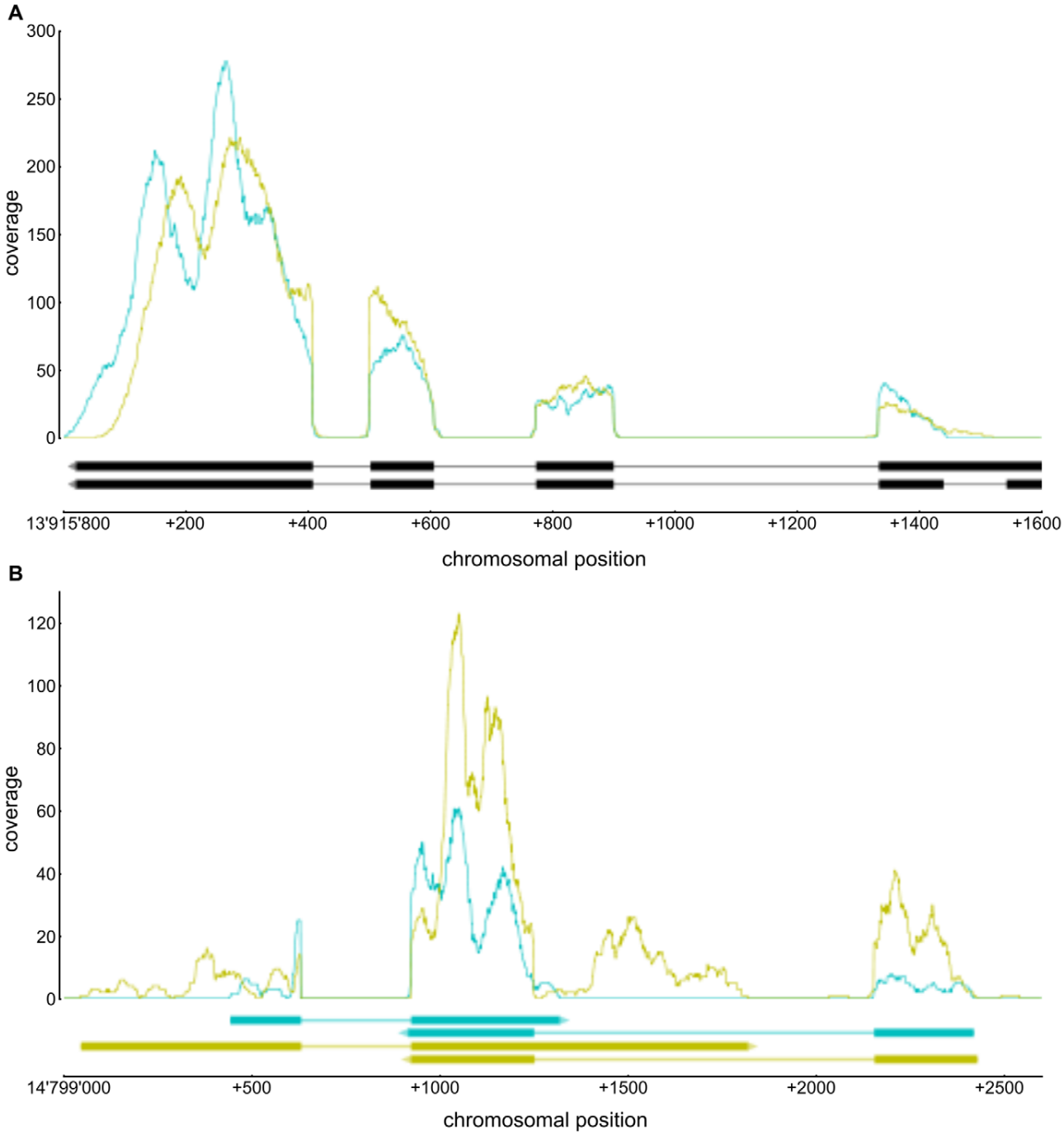


Figure 2. Examples of sequence coverage in annotated (A) and unannotated (B) regions. Graphs in the upper parts of the panels represent the number of hits per base within the two replicates (CC1: cyan, CC2: yellow). Transcripts are drawn in the lower parts of the panels: dark boxes represent exons, bright lines mark introns and the arrowhead depicts the direction of transcription. (A) Sequence coverage at the region around the locus *AT4G27960* (*UBC9*) on chromosome 4. The two transcripts represent two isoforms of *AT4G27960*. Clearly visible is the lack of coverage at the introns and the non-uniformity of sequence coverage with the maxima close to the 3′ end of the transcripts. (B) Sequence coverage at a region on chromosome 5, which is not annotated as being transcribed. Hits in this region were assembled into transcripts using cufflinks [11]. For each replicate, two transcripts with overlapping 3′ ends could be assembled (CC1: cyan, CC2: yellow). Notably, the sequence coverage along these transcripts resembles the coverage observed at annotated transcripts (A). Also visible are the unsharp transcript boundaries which vary between the replicates.
doi:10.1371/journal.pone.0029685.g002

and 16% with an overlap of at least 10%). Reasons for this may be a higher variability between the two replicates at low expression levels, a stronger influence of sequencing or alignment errors, and a sparse read coverage leading to a highly fragmented assembly. Splice-loci with higher expression values exhibited substantially higher overlaps, ranging from 17% (perfect overlap) to 48%

(overlap of at least 10%). However, the number of splice-loci with an overlap above a certain threshold increased substantially, when overlaps of splice-loci with loci from transcripts not supported by splice junctions were also considered to be valid (19% perfect overlap, 85% with an overlap of at least 10%), likely indicating a fragmented assembly due to a lack of gapped alignments. Given that the assembly, especially of transcripts with low to moderate expression levels, is strongly depending on sequencing depth [11], we expect that an increased sequencing depth together with the use of paired-end reads would improve the assembly and thus, the overlap between the replicates. Taken together, we suggest that the potentially new transcripts identified in this study with cufflinks should be considered as preliminary, still requiring further experimental exploration and validation. Nonetheless, we consider cufflinks as a valuable tool to start the search for potentially new transcripts in unannotated regions. It provides a basis to explore so far unknown transcribed regions also by other methods such as sequence alignment or gene prediction.

Transcriptional profiling. To get an overview of the hits mapping to annotated transcripts, we visualized the coverage at the transcripts (example in Figure 2A). This confirmed a 3′ bias, which was likely introduced during cDNA synthesis, within the data [15]. The 3′ bias partly counteracts the transcription length bias discussed in [20], due to a non-uniform coverage along a transcript. The relationship between the number of hits per transcript and its length is therefore only linear at the 3′ end of the transcript where the coverage is still uniform. Assuming a linear relationship over the entire length would thus lead to an underestimation of expression values from longer transcripts (e.g. RPKMs in the ERANGE software [21]). A possibility would be to take only hits in a certain distance to the 3′ end. However, this would exclude a certain proportion of the data [20]. We therefore decided to use the total number of hits mapping to the transcripts of a locus as expression value for the locus. Hits mapping to more than one locus (ambiguous hits) were proportionally distributed based on the number of unambiguous hits. Loci with transcripts having less than five hits or no hit located within the 250 bps at the 3′ end were discarded, the others declared as being expressed. Of the 33'598 annotated genes, pseudogenes, and transposable element genes, 17'419 (51.8%) met these criteria in at least one of the replicates (Table S2). Among these genes, 13'426 were found within both replicates. The other 3'993 loci were specific to one of the replicates (CC1: 1'028, CC2: 2'965). These loci had generally low expression values in the replicate in which they were detected (Figure 3A). It is therefore possible that a higher sequencing depth would reduce the number of replicate-specific loci. Beside this difference within the presumably low abundant transcripts, the replicates were highly similar (Figure 3A). However, the differences highlight the importance of replication that is necessary to estimate the variability in the data, especially of the genes with presumably low expression levels.

To compare the data generated with RNA-Seq to the one measured with the ATH1 microarrays [3], the expression values of the RNA-Seq data were transformed (log2(x+1)). ATH1 expression values and present calls were obtained as described in [22] (Table S2).

We first compared the average expression values. Using only the genes which have a corresponding probe set on the ATH1 array (21'440), we found that the measures of transcript abundance were well correlated (Figure 4A). Differences could be observed where array expression values were relatively high and the expression values from the RNA-Seq data small or zero (in agreement to [9]). This observation may be due to probe-specific background hybridization on the array [9].

We next compared the transcriptomes. From the 7'633 genes which were found to be expressed in the ATH1 array data, 93% were also detected in the RNA-Seq data (Figure 4B). The remaining 7%, exhibited medium expression values in the array data (Figure 4A). As mentioned before, it is possible that expression values for some of those genes were elevated due to probe-specific background hybridization. In addition to the 7'099 genes found in both data sets, 10'320 genes were only detected in the RNA-Seq data. From these, 34.6% were *a priori* not measurable using the ATH1 array because it lacks the corresponding probesets. The other 65.4% had low expression values in the array data. It is therefore likely that these signals could not be separated from the background [23].

Comparing RNA-Seq and ATH1 array data from central cells. Given the differences in the size of the transcriptomes, we investigated a potential effect on the functional characterization of

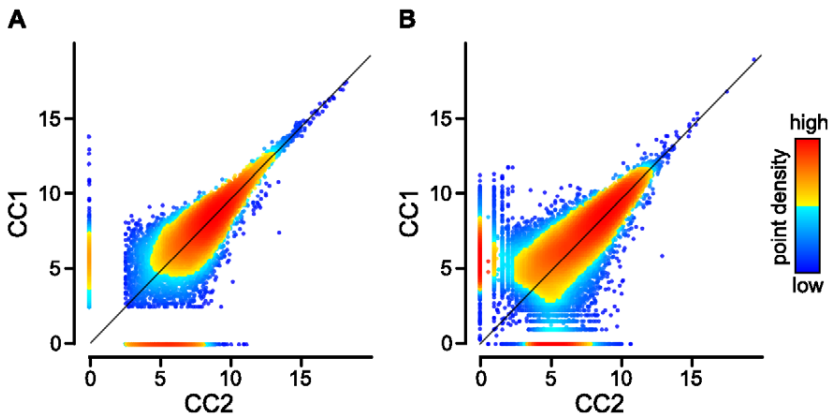


Figure 3. Comparisons of expression values between the two RNA-Seq replicates. In each panel, the expression values (log2 of the number of hits plus one) for each feature are plotted on the x-axis (CC2) and the y-axis (CC1). Colors indicate the point density: red and blue indicate the highest, respectively lowest, densities. (A) refers to the approach that was based on the alignment of reads to the reference genome: given are the expression values of the "expressed" genes (Pearson correlation: 0.99, Spearman correlation: 0.83). (B) refers to the approach that was based on *de novo* assembly of the short reads. Reads from both replicates were pooled and assembled together. To calculate expression values, reads from both replicates were aligned to the assembled transcriptome (Spearman correlation: 0.87).
doi:10.1371/journal.pone.0029685.g003

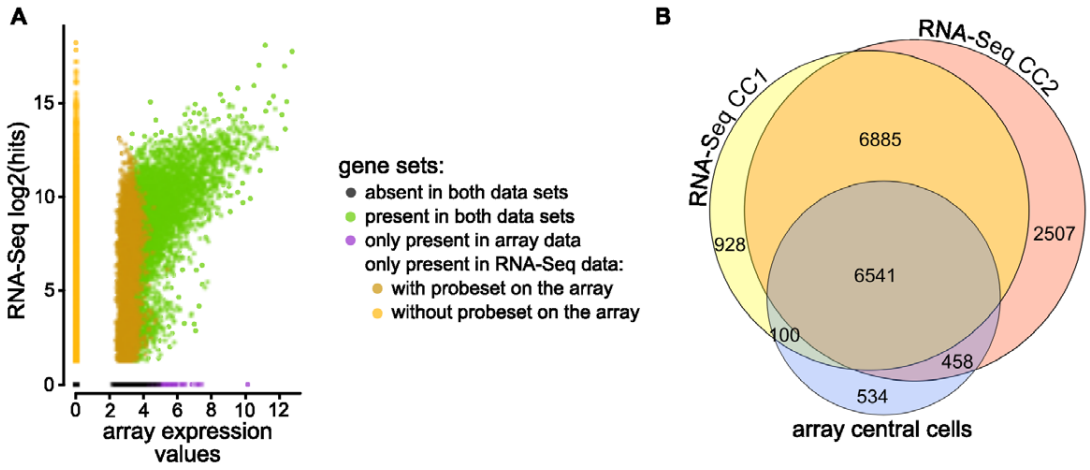


Figure 4. Comparisons between microarray and RNA-Seq data. (A) The average number of hits (log2(x+1)) for each gene are plotted on the y-axis and the corresponding normalized expression values from the array data are shown on the x-axis. Expression values of the genes having a probeset on the array are well correlated between the technologies (Spearman correlation: 0.63). (B) A Venn diagram summarizing the overlap between genes detected to be expressed in the RNA-Seq data sets and the array data.
doi:10.1371/journal.pone.0029685.g004

the central cell transcriptome. In terms of gene functions, the transcriptome measured with microarrays may either be a representative subset of the transcriptome determined using RNA-Seq or show a systematical over- or underrepresentation of genes having certain functions. Considering that the few array-specific genes were likely to be false positives, such systematic overrepresentation of functional groups in the array data would arise either from those false positives or from a sampling artifact, which may occur if a certain functional group was very well characterized at the time the ATH1 microarray was designed, and thus, almost entirely represented on the array. We therefore only tested for overrepresentation of transcripts encoding a given combination of protein domains (InterPro, www.ebi.ac.uk/interpro) in the RNA-Seq data compared to the array data (Figure 5, Table S3). Enrichment in the RNA-Seq data may originate from specific groups of genes newly detected due to either the higher sensitivity, which increases the propability to detect low expressed genes, or the whole-genome coverage that allows to detect genes which are not measurable with the ATH1 microarrays due to a lack of the corresponding probeset. The latter is a consequence of the ATH1 microarray design and would not occur with arrays covering the whole genome. We therefore performed two tests to separate the two effects from each other.

To test for enrichment likely caused by a higher sensitivity, we compared the functional characterizations of the transcriptomes determined using the array or the RNA-Seq data and excluded the genes missing a corresponding probeset on the ATH1 microarray. From 4'657 distinct (combinations of) protein domains found in this set of genes, 20 were significantly enriched in the RNA-Seq data compared to the array data (Fisher's exact test, one-sided p-value$<$0.05). Among them, (combinations of) domains which might play important roles in cell fate determination were identified: signal perception and transduction (Toll-like receptor), chromatin remodeling (SNF2-related helicase), regulation of transcription (SANT, Homeodomain-like, MYB), and RNA-binding (Figure 5).

To estimate the effect of the whole-genome coverage on the functional characterization, we performed a second enrichment analysis, which included also the genes missing a corresponding probeset on the ATH1 microarray and could identify seven additional (combinations of) protein domains showing significant enrichment in the RNA-Seq data. The largest group comprised genes with an "unknown" domain, which included uncharacterized protein-coding as well as non-protein-coding genes. The enrichment was therefore likely due to the non-protein-coding genes from which only few are represented on the ATH1 microarray. The other six groups contained genes encoding for domains of unknown function (DUF784, DUF239), meprin and tumour necrosis factor receptor associated factor homology domains (TRAF-like), F-box and F-box associated domains, S1 self-incompatibility related proteins (SI-S1-like), and small cysteine rich defensin-like proteins (SI-SLG-like/DEFL, Figure 5). Interesting to note is that the latter were implicated as signaling molecules required for pollen tube guidance in *Zea maize* and *Torenia fournieri* [24,25]. In *Arabidopsis thaliana* they might contribute to the role of the central cell in pollen tube guidance [3,26] or, what remains to be examined, as well function as signaling molecules within the mature embryo sac.

Taken together, we found that the two technologies correlate relatively well. Most of the transcripts detected in the array data were also detected in the RNA-Seq data (7'099). However, we could identify additional 10320 genes which are likely to be expressed in the central cell. A third of those could not be measured with the ATH1 microarray due to the lack of the corresponding probesets. This differences are therefore largely due to the ATH1 microarray design and would not occur in experiments using microarrays with whole-genome coverage. Importantly, however, the other two thirds could be attributed to the higher sensitivity of RNA-Seq compared to microarrays. Interestingly, this did not only increase the estimated size of the transcriptome, but also had an effect on the functional characterization of the transcriptome. Given that RNA-Seq is highly accurate [8,9,21,27], the results demonstrate the superior performance of RNA-Seq over the array based method for the transcriptional profiling of specific cell types. Nonetheless, at the moment microarrays still have certain advantages. Numerous tools were developed, tested, and used extensively for analysis of data from a broad range of experiments, offering reliable and efficient data analysis for almost any experiment. Compared to this, RNA-Seq data analysis is still a relatively new field of research which,

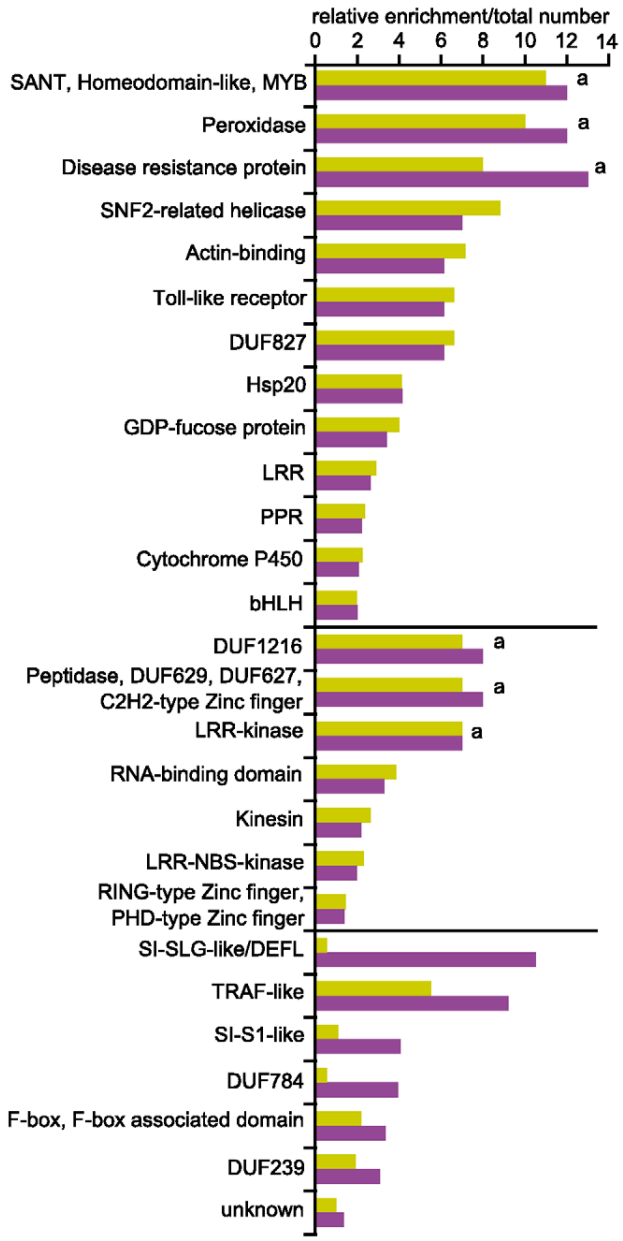


Figure 5. Test for enrichment of InterPro domains in RNA-Seq data compared to array data. The graph shows the relative enrichment of (combinations of) InterPro domains (simplified description, details are given in Table S3) in the RNA-Seq data compared to the array data, which was found to be significant. If the combination did not occur in the array data, the enrichment value was set to the total number of occurences of the combination in the RNA-Seq data (marked with a). We performed two tests to separate the effect of the higher sensitivity (yellow) from the effect caused by the whole-genome coverage (magenta). Combinations of protein domains in the upper, middle, and lower part of the figure were significantly enriched in both, the first, and the second test, respectively. Abbreviations: DUF: domain of unknown function, LRR: leucine rich repeat, PPR: pentatricopeptide repeat, bHLH: basic helix-loop-helix, NBS: nucleotide binding site, SI-: self-incompatibility, DEFL: defensin-like. The term "unknown" comprises all transcripts without an InterPro annotation (includes also non-protein-coding genes).
doi:10.1371/journal.pone.0029685.g005

however, evolves rapidly. Experience with the available tools is therefore rather limited compared to the ones used for microarray data analysis. Another advantage concerning the microarrays, which are frequently used, is the vast amount of publicly available data sets generated over the past years. For *Arabidopsis thaliana*, data from more than 7000 ATH1 microarrays are currently available on NCBI (www.ncbi.nlm.nih.gov). This offers the possibility to relate a newly determined transcriptome to many others, as for example presented in [3] where the transcriptomes of the cells from the female gametophyte could be directly compared to the ones of 59 different tissues or cell types. However, these advantages are likely to decrease fast as it is most probably only a matter of time until RNA-Seq will be the method of choice for transcriptional profiling [28].

Genes specifically expressed in central cells. A frequent application of transcriptional profiling is the analysis of differential expression of genes between different tissues and cell types or time points. With this approach, Wuest and coworkers [3] could identify 431 genes (FDR<0.05) which are likely to be specifically expressed in the mature female gametophyte as compared to 59 different tissues and cell types from *Arabidopsis thaliana*. Most of them were specific to one of the three cell types (113, 163, and 144 in the central cell, egg, and synergid cells, respectively). Functional characterization further revealed an enrichment of specific posttranscriptional regulatory modules and metabolic pathways in each cell type [3]. Given the higher sensitivity of RNA-Seq and the whole-genome coverage, we expect that an analysis using transcriptomes measured with RNA-Seq would provide an even deeper insight to the unique nature of the transcriptome of the mature female gametophyte. However, the small number of publicly available RNA-Seq data from *Arabidopsis thaliana* and the lack of RNA-Seq data from egg and synergid cells hamper an in-depth analysis as performed in [3]. Nonetheless, to get preliminary insights into the unique nature of the central cell transcriptome, we performed a comparison of the central cell transcriptome with transcriptomes from other tissues and cell types from *Arabidopsis thaliana*, which had been analyzed by RNA-Seq. The test set comprised data from 2–4 cell and globular stage embryos [12], early globular embryos [29], whole plants (pool of organs) [18], seedlings [18], unopened flower buds [19], and male meiocytes [30], thus representing a diverse set of tissues and cell types.

Using edgeR [31] to test for differential expression, we could identify 1'418 genes (FDR<0.05) upregulated in the central cell compared to the other tissues and cell types (Figure 6). We could thereby confirm 75% of the genes previously found to be specific to the central cell [3]. We also found that 9% and 17% of the genes previously described as enriched in the egg and the synergid cells, respectively, showed significant enrichment in the central cell in our comparison. Cross-contamination is however unlikely considering that the central cell is very well separatable from the egg and the synergids. In addition, one would rather expect contamination from the egg cell instead of the synergids, as the egg is closer to the central cell than the synergids. We therefore suggest that the difference was likely an artifact due to the lack of RNA-Seq data from the egg and synergid cells: In our comparison, genes expressed in central cells at a level below the one in egg or synergid cells but above the level in all other tissues were identified as specifically enriched. However, if data from egg and synergid cells were included, these genes would not be identified as being enriched in central cells.

To characterize the set of genes found to be specifically enriched in the central cell, we used the InterPro annotation (www.ebi.ac.uk/interpro) and tested for enrichment of certain (combinations of) protein domains as described above (Table S4). We found 118 and

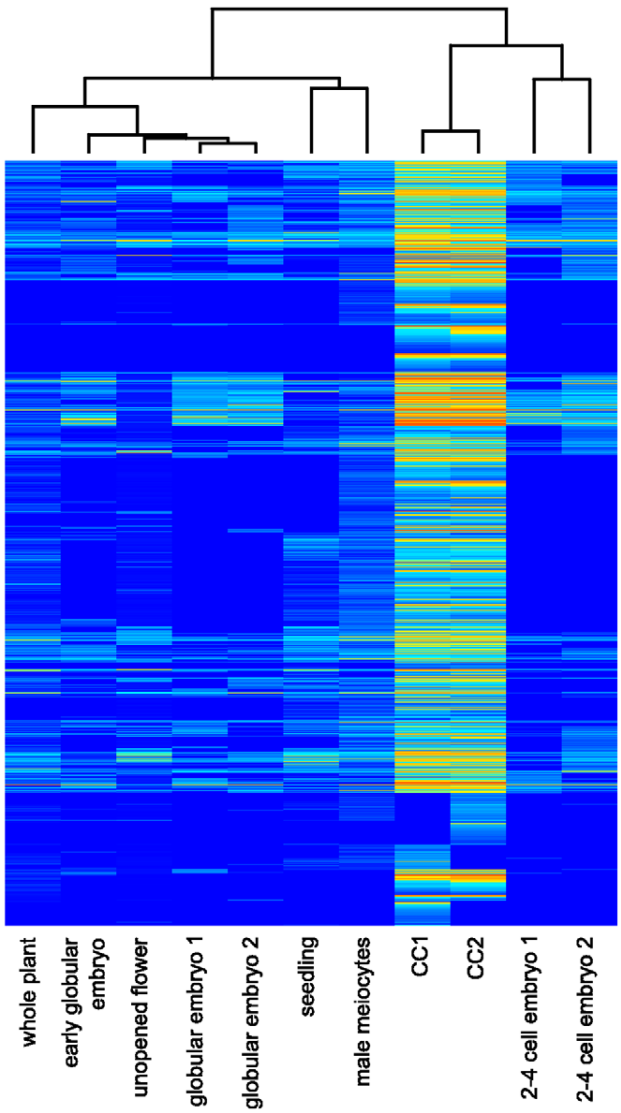


Figure 6. Genes enriched in the central cell compared to other tissues of *Arabidopsis thaliana*. Expression values of genes preferentially expressed in central cells are summarized in a heatmap (blue/red: low/high expression values). Expression values were equalized using edgeR [31] and log2(x+1) transformed. Samples and genes were clustered using Spearman correlation and hierarchical agglomerative clustering. Transcriptomes from whole plant and seedlings, unopened flowers, early globular embryos, male meiocytes, and 2–4 cell and globular stage embryos were obtained from [18], [19], [29], [30], and [12], respectively.
doi:10.1371/journal.pone.0029685.g006

11 (combinations of) protein domains showing enrichment in the central cell at a low stringency (p-value$<$0.05) and high stringency (FDR$<$0.05) cutoff, respectively. Among the most significantly enriched, are several domains that are underrepresented on the ATH1 array: domains of unknown function (DUF784, DUF239), F-box and F-box associated domains, S1 self-incompatibility related proteins, and small cysteine rich defensin-like proteins (DEFLs). These results indicate that genes specific to the mature female gametophyte are generally underrepresented on the ATH1 array as observed previously [32]. However, even though underrepresented on the ATH1 array, several of them (F-box, DUFs, DEFLs) were already identified previously to be specifically enriched in the mature female gametophyte [3], with the DEFL being highly specific to the central cell, thus indicating good agreement between the comparisons performed in this study and [3]. A similar overlap could also be observed for several (combinations of) protein domains which may play an important role in cell fate determination, such as the type I MADS-box and RWP-RK transcription factors. Examples for functional groups not identified in [3] comprise several genes encoding for diverse epigenetic functions including a histone methyltransferase (*AT2G24740/SUVH8*), a chromomethylase (*AT1G80740/CMT1*), and two DNA-methyltransferases (*AT4G08990* and *AT4G14140/MET2*), which may contribute to the specific epigenetic state of the central cell [33,34].

Taken together, the enrichment analysis presented in this study widely agrees with previously reported results [3] in terms of functional enrichment, but extends the number of specific genes to a large extent. However, given the few RNA-Seq data sets in the comparison and the lack of the egg and synergid transcriptomes, these results are preliminary, requiring additional data sets for the comparison in order to obtain a more thorough view on the unique nature of the transcriptome of the central cell.

Data analysis using *de novo* transcriptome assembly

RNA-Seq offers the possibility to investigate an organism which lacks sequence information. To test whether the methods presented here are suitable for such a study, we performed a *de novo* assembly of short reads and briefly characterized the transcriptome using GO terms. An in-depth analysis of the results is, however, out of scope of this article.

De novo assembly of RNA-Seq data into transcripts is generally challenging due to the non-uniform sequencing coverage across transcripts, the presence of low quality reads, and the size of the data sets [35]. In the data presented here, additional complexity is caused by fragment shortening introduced during library preparation, which leads to a non-uniform sequence coverage within the individual transcripts. To overcome some of the limitations, we first corrected potential sequencing errors and then removed all reads which were of low quality, repetitive or duplicated. The remaining reads were assembled using velvet [36], oases, and additional tools, which were required to handle the color space of SOLiD reads. Expression values were obtained using the full set of reads and bowtie [37]. To characterize the assembled transcripts, we used Blast2GO, which is designed to annotate (protein coding) sequences based on similarity searches and existing annotation associations [38]. Transcripts were first mapped against NCBI's non-redundant protein sequence database (www.ncbi.nlm.nih.gov) using BLAST [39] with an e-value threshold of 1e-6 and a maximum of 20 blast hits. Gene Ontology (GO) terms [40] were obtained and assigned using default settings.

From the initial set of reads containing reads from both replicates, around half (44'686'342) passed the filter criteria. From these, around 28.7% were used during the assembly, resulting in 32'567 transcripts with an average length of 307.1 bp (File S2) and a sequence coverage resembling the results from the previous analysis; beside the differences for replicate-specific transcripts, sequence coverage was highly similar (Figure 3B). From the 32'567 assembled transcripts, 19'502 had one or more blast hits to the non-redundant protein sequence database. Most (89.4%) of them had the best blast hit to *Arabidopsis thaliana* or its close relative *Arabidopsis lyrata*. In the latter case, the transcripts generally also mapped to *Arabidopsis thaliana* sequences. The majority of the remaining best blast hits were found among fungal pathogens affecting plants (3.8%) and animals (1.8%). Transcripts with hits to

fungal species originated from the first replicate (CC1) and were not found in the second replicate (CC2), indicating some contamination of the RNA from CC1 (replicate-specific assemblies, data not shown). Interestingly, reads aligning to those sequences do generally not align to the genome of *Arabidopsis thaliana* (0.0016% of all reads aligning to the genome do align to the sequences of those fungi). It is therefore unlikely that the contamination affected the alignment-based approach described before.

To compare the assembled transcriptome to the transcriptome determined in the alignment based approach, the two transcriptomes were annotated and compared at the level of GO terms using Blast2GO [38]. To avoid a bias due to the fungal contamination, we only used the assembled transcripts with the best blast hit to either *Arabidopsis thaliana* or *Arabidopsis lyrata* for the comparison. From these 17'641 transcripts, 14'514 could be annotated with 4'859 distinct GO terms. Overall, the number of distinct GO terms per annotated transcript was lower in the transcriptome from the *de novo* assembly compared to the transcriptome determined with the alignment based approach described before, where 14'487 of the 17'419 transcripts could be annotated with 5'285 distinct GO terms (only one, the representative, transcript per locus). However, the distribution of GO terms belonging to the domain "Molecular Function" showed a high similarity between the two transcriptomes: only 10 terms showed significant enrichment in one of the transcriptomes (two-sided Fisher's exact test, FDR$<$0.05). The most specific terms among them were: structural constituent of ribosome (GO:0003735), transcription factor activity (GO:0003700), RNA binding (GO:0003723), protein serine/threonine kinase activity (GO:0004674), and translation factor activity/nucleic acid binding (GO:0008135). All those terms displayed an enrichment in the assembled transcriptome. For genes related to transcription factor or protein kinase activity this was unexpected as they are often expressed at low levels. However, it is possible that the marginal coverage of these transcripts caused a fragmented assembly: if the reads from one transcript were assembled into two fragments of the transcript, from which both map to the GO term, the GO term would be counted twice, thus leading to an overrepresentation of the respective GO term.

Taken together, the results of this test indicate that data obtained with the methods presented here is in principle suitable for *de novo* assembly of a transcriptome: sequence coverage patterns and GO annotations largely resembled the ones found in the alignment-based approach. However, it is likely that many of the assembled transcripts were shorter than *in vivo* due to the 3′ bias. In cases where most of the assembled sequence contained mainly untranslated regions (long 3′ UTR), it probably had an effect on the GO term annotation (which is based on similarity to known proteins). Considering further that the annotation using GO terms can only characterize protein-coding transcripts, we recommend to use additional methods for the annotation and analysis of the assembled transcripts. One possibility would be to search databases containing all types of transcripts for similarity in the nucleotide sequence. We expect that this would help to characterize non-coding transcripts and improve the GO annotation of protein-coding transcripts which could not be annotated using the methods relying on similarity to proteins.

Conclusion

We aimed to establish a workflow that allows determining the transcriptional profile with a high sensitivity and resolution of specific cell types, which are very rare and difficult to access as they are embedded in heterogenic tissue. We therefore combined LAM with a highly sensitive, linear RNA amplification method and the emerging RNA-Seq technology. As a model we used central cells of *Arabidopsis thaliana* from which only around 50 are formed within a flower, each of them individually enclosed by an ovule. Using LAM, we could obtain sufficient amounts of good quality RNA for a successful amplification and library preparation. We compared the data generated in this study with the transcriptome data from [3], which was measured using LAM and the ATH1 microarray. The results showed that the two transcriptome profiling technologies correlate well. Most of the genes found to be expressed in the microarray data were also present in the RNA-Seq data and the few microarray specific genes were likely false positives caused by probe specific cross-hybridization. However, using RNA-Seq we could detect more than double the amount of presumably expressed genes. Functionally, this difference was reflected in the enrichment of genes encoding for few specific (combinations of) protein domains, of which some may play an important role in cell fate determination (signal perception and transduction, chromatin remodeling, and regulation of transcription) or function of the specific cell type (defensin-like proteins), in the RNA-Seq data compared to the array data. In addition, we identified several intergenic regions which are likely to be transcribed. We further described a considerable fraction of reads aligning to introns and regions flanking annotated loci which may represent alternative transcript isoforms. Finally, we also performed a *de novo* assembly of short reads and briefly characterized the assembled transcriptome. Comparisons between the alignment- and the assembly-based approaches revealed that the results were remarkably similar in terms of sequence coverage pattern and Gene Ontology (GO) annotation, indicating that the workflow presented here is also suitable to study specific cell types from an organism lacking a reference sequence. Taken together, we successfully established an easy and reliable workflow that allows the transcriptional profiling of specific cell types, which are rare and difficult to access, with high sensitivity and resolution. The approach presented here will provide new insights into the transcriptional state of individual cell types not only of plants, but also other eukaryotes and, therefore, by elucidating cell fate decisions, will contribute to the understanding of the molecular processes underlying the development of multicellular organisms.

Materials and Methods

Plant material

Arabidopsis thaliana seeds, accession Landsberg *erecta*, were germinated on Murashige and Skoog agar (0.5× Murashige and Skoog salts, 0.7% phytagar). One week old seedlings were transferred to ED73 soil (Einheitserde, Schopfheim, Germany), and grown under 16 h light at 21°C and 8 h darkness at 18°C and 60% relative humidity.

Tissue embedding

Two days after emasculation, flowers containing the mature embryo sacs were fixed in ethanol:acetic acid 3:1. Vacuum was applied two times for 15 min at 4°C. Afterwards the material was kept in the fixative overnight at 4°C and subsequently transferred to an ASP200 embedding machine (Leica Microsystems GmbH, Wetzlar, Germany). In the embedding machine, tissues were dehydrated automatically in a graded series of ethanol (70% for 1 h, 3×90% for 1 h, 3×99.98% for 1 h, all at room temperature) followed by xylol (2×1 h and 1×1 h 15 min, all at room temperature). Xylol was substituted by Paraplast X-tra embedding media (Roth AG, Arlesheim, Switzerland) at 58°C (2×1 h and

1×3 h). Finally, flowers were poured into paraffin blocks, cooled, and stored at 4°C.

Laser-assisted microdissection

For microdissection, paraffin blocks containing flowers were cut on a RM2145 Leica microtome (Leica Microsystems GmbH, Wetzlar, Germany) to 8 µm thin sections and mounted on nuclease-free membrane-mounted metal-frame slides using pure methanol ([3] used water). Slides were dried overnight on a heating table at 42°C. Samples were deparaffinized in xylol at room temperature (2×10 min). Microdissection was performed using a mmi CellCut Plus device (MMI Molecular Machines & Industries AG, Glattbrugg, Switzerland). Isolated central cells were collected using MMI isolation caps and stored at −80°C. Four to six cuts of whole flowers were taken from each slide after collecting the cells of interest. Total RNA was isolated and tested for integrity using the Agilent 2100 Bioanalyzer together with the RNA 6000 Pico Kit (Agilent Technologies Schweiz AG, Basel, Switzerland).

RNA isolation

Total RNA was isolated using the PicoPure RNA isolation kit (Arcturus Engineering, Mountain View, USA) following the manufacturer's instructions with slight modification. Caps were covered with 10 µl extraction buffer and incubated at 42°C for 30 minutes. Extracts from different caps were pooled to reach a sufficient RNA yield. All other steps were performed according to the manufacturer's instructions, including the on-column DNase treatment using the RNase-free DNase set from QIAGEN (Valencia, USA).

RNA sequencing

The amplification and library preparation of RNA from central cell samples were performed following the protocol described in [14]. Libraries were sequenced using the SOLiD platform, version 3 (Life Technology, Foster City, USA). For each library one eighth of a slide was used.

qRT-PCR

To monitor the efficiency of the library preparation we measured the size of the cDNAs prior to shearing using Agilent's 2100 Bioanalyzer following the instructions from the manufacturer. We also estimated the concentration of cDNA from selected genes with qRT-PCR: *ACT2* (*AT3G18780*, forward: CTTGCACCAAGCAGCATGAA, reverse: CCGATCCAGACACTGTACTTCCTT, [41]), *ACT11* (*AT3G12110*, forward: AAGCTGTTCTTTCCCTCTACGC, reverse: GGAACAGTGTGACTCACACCATC, [42]), *EF*−1α (*AT5G60390*, forward: TGAGCACGCTCTTCTTGCTTTCA, reverse: GGTGGTGGCATCCATCTTGTTACA, [43]) and *UBC9* (*AT4G27960*, forward: TCACAATTTCCAAGGTGCTGC, reverse: TCATCTGGGTTTGGATCCGT, [41]). All qRT-PCR reactions were performed in a final volume of 10 µl containing 5 µl cDNA or water, 0.125 µl of each primer (5 µM), 0.25 µl water and 4.5 µl mastermix (Power SYBR Green PCR Master Mix, Life Technology) on the 7900 HT Fast Real Time PCR System (Life Technology) with the following cycling conditions: 50°C for 2 minutes, 95°C for 10 minutes and 45 times 95°C for 15 seconds followed by 60°C for 1 minute. Amplicon length was determined using the melting curve analysis.

Data processing: reference genome

Short reads and alignments generated in this study were deposited at NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) and are accessible through GEO series accession number GSE29719. Reads (csfasta and qual files) were processed with the SOLiD Accuracy Enhancement Tool (version 2.2 with a reflength of 13'000'000 and the option -qvupdate; solidsoftwaretools.com/gt/project/saet [note added in proof: SAET was moved to solidsoftwaretools.com/gf/project/denovo/frs]) and aligned to the reference genome (www.arabidopsis.org) using TopHat with the options –color –quals –coverage-search -a 8 -m 1 -i 50 -I 2000 -F 0.2 -p 7 -g 10 (version 1.2; [11]). We allowed up to ten alignments per read to avoid a potential underestimation of expression values of transcripts with similar sequence. However, a read r with $m>1$ alignments would count m times, resulting in overestimation of expression values. To avoid this, we calculated for each alignment i of such a read the weight H_i using a "score" S_i divided by the sum of scores from all alignments of the read ($H_i = S_i / \sum_{i=1}^{m} S_i$). If the total score was zero, all alignments were discarded. For ungapped alignments, the score was equal to the sum of coverage originating from uniquely aligned reads at the position of the alignment and the surrounding 100 bps ("allocation distance" of ±50 b*ps*). For gapped alignments, the score was equal to the number of uniquely aligned reads spanning the same gap. Thus, if a read had both types of alignments, the ungapped ones would have been preferred. Here we use "hit" as a synonym for an alignment that has been weighted.

Identification of new transcripts. To find potentially new transcripts in intergenic regions, we extracted all alignments that were not overlapping with a known transposable element and at least 200 bps outside of a known locus. The "intergenic" transcriptome was then assembled using these intergenic alignments and cufflinks (version 0.9.3) with a maximal intron length of 2000 [11]. To compare the genomic location of the loci from the potentially new transcripts between the two replicates, we calculated for each locus from each replicate the overlap with a locus/loci (with and without the remaining loci with transcripts not supported by splice junctions) from the other replicate (number of shared nucleotides divided by the length of the locus) and counted the number of loci with an overlap above a certain threshold. Potentially new transcripts supported by at least one splice junction were annotated using Blast2GO (version 2.4.8; [38]).

Transcriptome data. Hits were assigned to the transcripts of the genomic features "gene", "pseudogene" and "transposable element gene" (TAIR10, www.arabidopsis.org). Hits can be divided into unambiguous (mapping to transcripts of only one locus) and ambiguous (mapping to transcripts of more than one locus). To avoid counting ambiguous hits multiple times, we proportionally distributed them based on the number of unambiguous hits. If there were no unambiguous hits, the ambiguous hits were equally distributed. However, we assume a case where two loci A and B overlap such that locus A is entirely located within locus B. Locus A shall be "truly" expressed, locus B not. Using a single step, all hits of locus A would be declared as ambiguous. In case locus B has no unambiguous hit, the hits from locus A would be equally distributed to locus A and B, leading to an underestimation of the expression value from locus A and an overestimation of the expression value from locus B (a false positive). In another case where locus B has one or two unambiguous hits due to sequencing and/or alignment errors, all the hits from locus A would be wrongly assigned to locus B (one false positive and one false negative). The same error would occur if locus A has a longer transcript than the annotation would indicate. The hits at the borders of locus A would then be unambiguously assigned to locus B and as a consequence also all the ambiguous hits. To avoid this scenario at least to some extent we used a two step approach. In the first step, all hits were mapped to all annotated transcripts. We expected that each "truly"

expressed transcript should have at least one hit within the 250 bps at its 3′ end because the library preparation protocol was based on poly(A)-tail priming for cDNA synthesis and adapters for the amplification. In addition, we set a threshold of five hits as a minimal expression value to overcome possible sequencing and alignment errors. Transcripts not matching these criteria were discarded. During the second step, the hits were divided into unambiguous and ambiguous. The unambiguous hits were assigned first and used to distribute the ambiguous hits. The transcripts were then filtered again using the same criteria as before. The final expression value of a locus was calculated as the sum of hits assigned to any of its transcripts. Expression values are given in Table S2.

Enrichment of combinations of protein domains (InterPro). Genes present in array data (i), RNA-Seq data (ii), and RNA-Seq data excluding genes lacking a corresponding probeset on the array (iii) were functionally characterized using the InterPro annotation (www.ebi.ac.uk/interpro). Information necessary to map the InterPro terms to the *Arabidopsis thaliana* gene identifiers was extracted from the functional gene descriptions available on www.arabidopsis.org (genes with no annotation were annotated as "unknown"). Some terms in the InterPro annotation are hierarchically linked to each other. Given this "parent to child" relation, a gene annotated with one term is automatically also annotated with all the ancestors of the term. To avoid reduction of statistical power due to this dependencies, we only used the lowest possible terms to characterize the genes. All terms annotating a gene were then grouped together, forming a specific combination of protein domains. To test for enrichment of a given combination in the RNA-Seq data, occurences were calculated and compared using Fisher's exact test (one-sided). Combinations with a p-value below 0.05 were declared to be significantly enriched (due to redundancies in the InterPro annotation, multiple testing correction may have been to stringent).

Genes preferentially expressed in central cells. The transcriptome of the central cell was compared to publicly available RNA-Seq transcriptome data from various tissues and cell types of *Arabidopsis thaliana*. The data comprised 2–4 cell and globular stage embryos [12] (GSE24198, GSE33866), early globular embryos [29] (SRR074122), whole plants (pool of organs) [18] (SRR018346, SRR018347, SRR019035), seedlings [18] (SRX006704), unopened flower buds [19] (SRX002554), and male meiocytes [30] (SRX063784). Raw data (csfasta/qual and fastq files) were downloaded from www.ncbi.nlm.nih.gov/geo/ (GSE accession numbers) and trace.ddbj.nig.ac.jp/DRASearch/ (SRX/SRR accession numbers). Only data from untreated wild-type plants were used in the analysis. The data was largely processed as described above with modifications depending on the experimental setup and without the thresholds of 5 hits and at least one hit in the first 250 bp of a transcript. In the data sets from [12] (50 bps reads, SOLiD), reads with multiple alignments were removed due to their high abundance (a consequence of the amplification strategy using random hexamers in addition to the poly(T)-primers for cDNA synthesis). For the remaining data sets from [18,19,29,30] (36 bps reads, Illumina), reads could not be corrected and the allocation distance was set to ± 36 b*ps*. Genes preferentially expressed in central cells compared to all other tissues and cell types were then identified with edgeR [31] using tagwise dispersion estimates and Benjamini-Hochberg multiple testing corrections. Genes with an adjusted p-value (FDR) below 0.05 were considered to be differentially expressed. To test for enrichment of certain (combinations of) protein domains in the central cell transcriptome, we compared the functional characterization of the genes significantly upregulated in central cells with the one of the genes showing no significant differential expression using the approach described above (Table S4).

Data processing: *de novo* assembly

Reads were corrected as described above. We removed all reads which were of low quality (total quality below 200 or an ambiguous color in the sequence), repetitive (same double color in more than 30% of the sequence), or duplicated. Assembly was performed on double encoded reads using velvet (version 1.0.18 [36]) and oases (version 0.1.18, www.ebi.ac.uk/~zerbino/oases) with a k-mer length of 31, a minimal transcript length of 80, and a minimal coverage of 1. Double encoding and decoding was done using the pre- and postprocessor scripts (versions 2.2.1 and 1.6, solidsoftwaretools.com/gt/project/denovotools) in conjunction with asid_light (version 1.0, solidsoftwaretools.com/gt/project/denovo). All reads were then mapped back to these assembled reference transcriptomes using bowtie with the options -C -n 2 -l 25 -k 11 -m 10 –chunkmbs 1024 –best –strata -p 7 (version 0.12.7; [37]). Ambiguous alignments were proportionally distributed using the number of unambiguous alignments. The final expression values were calculated as the sum of hits mapping to a transcript. Assembled transcripts and representative gene models from the reference annotation (www.arabidopsis.org) were annotated using Blast2GO (version 2.4.8; [38]). For blastx against the non-redundant protein sequences deposited at NCBI (www.ncbi.nlm.nih.gov) an e-value threshold of 1e-6 was chosen. Parameters for the GO annotation and analysis were left at default. To compare the annotations, we used the tool embedded in Blast2GO (Blast2GO version 2.4.9). GO terms with an FDR below 0.05 were defined as being significantly differentially enriched (two-sided Fisher's exact test).

Microarray data

Microarray data [3] were obtained from ArrayExpress (www.ebi.ac.uk/arrayexpress, accession number E-MEXP-2227) and processed as described in [22]. Final expression values are given in Table S2.

Software

Unless specified, we used newly developed software. The core package is split into several programs which are largely independent of each other (processing of reads with multiple alignments, filtering of genes, distribution of ambiguous hits, filter for *de novo* assembly) and therefore offers flexibility to customize and extend the analysis. Source code and linux binaries for the transcriptome analysis are freely available upon request (schmid.m@access.uzh.ch).

MIAME

All data are MIAME compliant. Raw data were deposited at (RNA-Seq data: GSE29719 on GEO) and obtained from (microarray data: E-MEXP-2227 on ArrayExpress; RNA-Seq data: GSE24198, GSE33866, SRR074122, SRR018346, SRR018347, SRR019035, SRX006704, SRX002554, SRX063784 on GEO and DRASearch) MIAME compliant databases.

Supporting Information

File S1 The file contains the results from the cDNA library control experiments (size distribution of fragments and approximate concentration of selected genes).
(PDF)

File S2 The rar file contains the transcript sequences from the *de novo* assembly (fasta file).
(RAR)

Figure S1 The figure contains a summary of the overlaps of the potentially new loci producing transcripts supported by splice junctions given in Table S1 between the two replicates.
(PDF)

Table S1 The table contains the genomic coordinates and annotations of the potentially new transcripts, which were identified and annotated using cufflinks and Blast2GO, respectively. Only transcripts supported by at least one splice junction are presented.
(XLS)

Table S2 The table contains the RNA-Seq expression values from all genes declared to be present in at least one of the replicates (sheet 1) and the microarray expression values from all the genes having a corresponding probeset on the microarray (sheet 2).
(XLS)

Table S3 The table contains the results from tests for enrichment of InterPro domains in the RNA-Seq data compared to the array data. In the first test (sheet 1), the gene universe was defined as the union of all the genes found to be expressed in any of the data type. In the second test (sheet 2), genes, which were present in the RNA-Seq data but are per default not detectable with the array due to the lack of a corresponding probeset, were excluded from the universe. The third sheet contains a table with additional information to Figure 5.
(XLS)

Table S4 The table contains the results from tests for enrichment of InterPro domains in the central cell transcriptome compared to transcriptomes of other tissues. The test set contained all genes showing significant enrichment in the central cell. The reference set comprised all the other genes (only the ones sequenced in at least one experiment).
(XLS)

Acknowledgments

We thank Samuel E. Wüst (Trinity College Dublin) for helpful discussions. We also acknowledge Célia Baroux (University of Zürich), Michael T. Raissig (University of Zürich), and Michael Wittig (Christian Albrechts University Kiel) for providing the data of the GSE33866 series.

Author Contributions

Conceived and designed the experiments: MWS AS UG. Performed the experiments: MWS UCK. Analyzed the data: MWS MB. Wrote the paper: MWS. Helped draft the manuscript: AS UG. Participated in design of analysis: PR. Revised and approved the manuscript: MWS AS UCK MB PR UG.

References

1. Kamme F, Salunga RC, Yu J, Tran D, Zhu J, et al. (2003) Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogenity. Journal of Neuroscience 23: 3607–3615.
2. Luo L, Salunga RC, Guo H, Bittner A, Joy KC, et al. (1999) Gene expression profiles of lasercaptured adjacent neuronal subtypes. Nature Medicine 5: 117–122.
3. Wüst SE, Vijverberg K, Schmidt A, Weiss M, Gheyselinck J, et al. (2010) Arabidopsis female gametophyte gene expression map reveals similarities between plant and animal gametes. Current Biology 20: 1–7.
4. Morse AM, Carballo V, Baldwin DA, Taylor CG, McIntrye LM (2010) Comparison between Nu-GEN's WT-Ovation Pico and One-Direct Amplification Systems. Journal of Biomolecular Techniques 21: 141–147.
5. Brooks III L, Strable J, Zhang X, Ohtsu K, Zhou R, et al. (2009) Microdissection of shoot meristem functional domains. PLoS Genetics 5: e1000476.
6. Day RC, McNoe L, Macknight RC (2007) Evaluation of global RNA amplification and its use for high-throughput transcript analysis of laser-microdissected endosperm. International Journal of Plant Genomics. 61028 p.
7. Kerk NM, Ceserani T, Tausta SL, Sussex IM, Nelson TM (2003) Laser capture microdissection of cells from plant tissues. Plant Physiology 132: 27–35.
8. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63.
9. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research 18: 1509–1517.
10. Garg R, Patel RK, Tyagi AK, Jain M (2011) *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Research 18: 53–63.
11. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28: 511–515.
12. Autran D, Baroux C, Raissig MT, Lenormand T, Wittig M, et al. (2011) Maternal epigenetic pathways control parental contributions to *Arabidopsis* early embryogenesis. Cell 145: 707–719.
13. Schneitz K, Grossniklaus U (1998) The molecular and genetic basis of ovule and megagametophyte development. Seminars in Cell and Developmental Biology 9: 227–238.
14. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods 6: 377–382.
15. Klostermeier U, Barann M, Wittig M, Hasler R, Franke A, et al. (2011) A tissue-specific landscape of sense/antisense transcription in the mouse intestine. BMC genomics 12: 305.
16. Tariq M, Kim H, Jejelowo O, Pourmand N (2011) Whole-transcriptome RNAseq analysis from minute amount of total RNA. Nucleic Acids Research 39: e120.
17. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111.
18. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, et al. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Research 20: 45–58.
19. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BG, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133: 523–536.
20. Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-seq data confounds systems biology. Biology Direct 4: 14.
21. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods 5: 621–628.
22. Schmidt A, Wüst SE, Vijverberg K, Baroux C, Kleen D, et al. (2011) Transcriptome analysis of the *Arabidopsis* megaspore mother cell uncovers the importance of RNA helicases for plant germ line development. PLoS Biology 9: e1001155.
23. Gentleman R, Huber W, Carey VJ, Irizarry RA, Dudoit S (2005) Bioinformatics and computational biology solutions using R and Bioconductor. New York: Springer Science+Business Media, LLC, first edition.
24. Márton M, Cordts S, Broadhvest J, Dresselhaus T (2005) Micropylar pollen tube guidance by Egg Apparatus 1 of maize. Science 307: 573–576.
25. Okuda S, Tsutsui H, Shiina K, Sprunck S, Takeuchi H, et al. (2009) Defensin-like polypeptide LUREs are pollen tube attractants secreted from synergid cells. Nature 458: 357–361.
26. Chen Y, Li H, Shi D, Yuan L, Liu J, et al. (2007) The central cell plays a critical role in pollen tube guidance in *Arabidopsis*. The Plant Cell Online 19: 3563–3577.
27. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344–1349.
28. Costa C, Angelini C, De Feis I, Ciccodicola A (2010) Uncovering the complexity of transcriptomes with RNA-Seq. Journal of Biomedicine and Biotechnology. 853916 p.
29. Nodine MD, Bartel DP (2010) MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. Genes & Development 24: 2678–2692.
30. Chen C, Farmer AD, Langley RJ, Mudge J, Crow JA, et al. (2010) Meiosis-specific gene discovery in plants: RNA-Seq applied to isolated *Arabidopsis* male meiocytes. BMC Plant Biology 10: 280.
31. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11: R25.
32. Jones-Rhoades MW, Borevitz JO, Preuss D (2007) Genome-wide expression profiling of the *Arabidopsis* female gametophyte identifies families of small, secreted proteins. PLoS Genetics 3: 1848–1861.
33. Hsieh TF, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, et al. (2009) Genome-wide demethylation of *Arabidopsis* endosperm. Science 324: 1451–1454.

34. Gehring M, Bubb KL, Henikoff S (2009) Extensive demethylation of repetitive elements during seed development underlies gene imprinting. Science 324: 1447–1451.
35. Martin J, Bruno VM, Fang Z, Meng X, Blow M, et al. (2010) Rnnotator: an automated *de novo* transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics 11: 663.
36. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18: 821–829.
37. Langmead B, Trapnell C, M P, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10: R25.
38. Conesa A, S G, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.
39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.
40. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25–29.
41. Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible WR (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. Plant Physiology 139: 5–17.
42. Baroux U, Gagliardini V, Page DR, Grossniklaus U (2006) Dynamic regulatory interactions of Polycomb group genes: MEDEA autoregulation is required for imprinted gene expression in *Arabidopsis*. Genes & Development 20: 1081–1086.
43. Czechowski T, Bari RP, Stitt M, Scheible WR, Udvardi MK (2004) Real-time RT-PCR profiling of over 1400 *Arabidopsis* transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. The Plant Journal 38: 366–379.
44. Christensen CA, King EJ, Jordan JR, Drews GN (1997) Megagametogenesis in *Arabidopsis* wild type and the Gf mutant. Sexual Plant Reproduction 10: 49–64.

5 Rcount: simple and flexible RNA-Seq read counting

The following manuscript is published in "Bioinformatics" (published by Oxford University Press, all rights reserved)[1]. I designed and implemented Rcount, wrote the manuscript and the user guide, and created all figures. Diana E. Coman Schmid critically read the manuscript, tested the software, and thereby provided valuable feedback to improve the manuscript and the user guide. Ueli Grossniklaus read and corrected the final draft and contributed the final name of the software.

[1] Schmid, MW and Grossniklaus, U (2015) Rcount: simple and flexible RNA-Seq read counting. Bioinformatics 31: 436–437.

BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 3 2015, pages 436–437
doi:10.1093/bioinformatics/btu680

Gene expression

Advance Access publication October 15, 2014

Rcount: simple and flexible RNA-Seq read counting

Marc W. Schmid* and Ueli Grossniklaus

Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zurich, 8008 Zürich, Switzerland

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: Analysis of differential gene expression by RNA sequencing (RNA-Seq) is frequently done using feature counts, i.e. the number of reads mapping to a gene. However, commonly used count algorithms (e.g. HTSeq) do not address the problem of reads aligning with multiple locations in the genome (multireads) or reads aligning with positions where two or more genes overlap (ambiguous reads). Rcount specifically addresses these issues. Furthermore, Rcount allows the user to assign priorities to certain feature types (e.g. higher priority for protein-coding genes compared to rRNA-coding genes) or to add flanking regions.

Availability and implementation: Rcount provides a fast and easy-to-use graphical user interface requiring no command line or programming skills. It is implemented in C++ using the SeqAn (www.seqan.de) and the Qt libraries (qt-project.org). Source code and 64 bit binaries for (Ubuntu) Linux, Windows (7) and MacOSX are released under the GPLv3 license and are freely available on github.com/MWSchmid/Rcount.

Contact: marcschmid@gmx.ch

Supplementary information: Test data, genome annotation files, useful Python and R scripts and a step-by-step user guide (including run-time and memory usage tests) are available on github.com/MWSchmid/Rcount.

Received on July 15, 2014; revised on September 24, 2014; accepted on October 13, 2014

1 INTRODUCTION

RNA-Seq is the method of choice for transcriptional profiling and differential expression (DE) studies. For DE analysis, methods based on negative binomial modeling, such as the popular DESeq (Anders *et al.*, 2010) and edgeR (Robinson *et al.*, 2010), have been shown to outperform other methods in terms of specificity, sensitivity and control of false positives (Rapaport *et al.*, 2013). Current work flows for DE analysis generally involve the (i) alignment of the short reads to a reference genome, (ii) quantification of expression levels and (iii) comparison between different treatments, tissue/cell types and time-points (Anders *et al.*, 2013).

Read counting and read summarization are essential steps in any RNA-Seq workflow. However, they have received little attention. Specifically for RNA-Seq, counting is not as simple as it may appear. First, a read may align multiple times with the genome (multireads). Second, several genes may overlap at a given position within the genome. Reads aligning with those positions are ambiguous with respect to the gene they originate from (ambiguous reads). Third, alignments can span exon-junctions (exon-junction reads). Furthermore, a gene may have several isoforms. However, DE analysis is often performed using the total number of reads per gene.

Popular open source tools for read counting, such as HTSeq (www-huber.embl.de/users/anders/HTSeq), BEDTools (Quinlan and Hall, 2010) and featureCounts (Liao *et al.*, 2014), do not specifically address all three issues. Multireads are not treated specifically by any of these programs and are generally discarded, although this problem has been addressed for ChIP-Seq data analysis (Chung *et al.*, 2011). Ambiguous reads are counted repeatedly for each gene by BEDTools and featureCounts (optionally), whereas HTSeq discards them. HTSeq and featureCounts both consider exon-junction reads, whereas BEDTools does not. ERANGE addresses all three problems, but uses RPKM (reads per kilobase per million) instead of counts throughout the whole algorithm. Moreover, a conversion to counts during the algorithm is not possible (Mortazavi *et al.*, 2008).

Here we describe Rcount, a fast and simple GUI tool for flexible RNA-Seq read counting. It builds on the algorithm described in Schmid *et al.* (2012), and further allows for editing the genome annotation and assigning priorities to certain feature types (see Figure 1A for details on genomic feature types).

2 DESCRIPTION

Rcount takes read alignments files (BAM, Binary Alignment/Map) and a reference genome annotation (GFF/GTF/BED, General Feature Format/Gene Transfer Format/Browser Extensible Data) as input, and counts the number of reads per gene, taking into account multireads, ambiguous reads and exon-junction reads (Fig. 1). It has three modules: *Rcount-multireads*, *Rcount-format* and *Rcount-distribute*.

Rcount-multireads assigns weights to each alignment of a multiread (Fig. 1B). The weight H_i of an individual alignment i is calculated using a score S_i divided by the sum of scores from all alignments of the multiread ($H_i = S_i / \sum_{i=1}^{m} S_i$). S_i is currently implemented as the sum of coverage (number of reads per base) originating from uniquely aligned reads at the position of the alignment i and the surrounding region (the size can be set by the user). If an alignment spans an exon junction, S_i equals to the number of uniquely aligned reads spanning the same exon junction. Thus, if a multiread has both types of alignments, the ungapped ones are generally preferred. The weight is automatically added as XW:f:H_i tag to the alignments in the BAM file.

Rcount-format reads the reference genome annotation in GFF/GTF/BED format, displays the structure of the genome annotation and saves it in an XML format required by *Rcount-distribute*.

*To whom correspondence should be addressed.

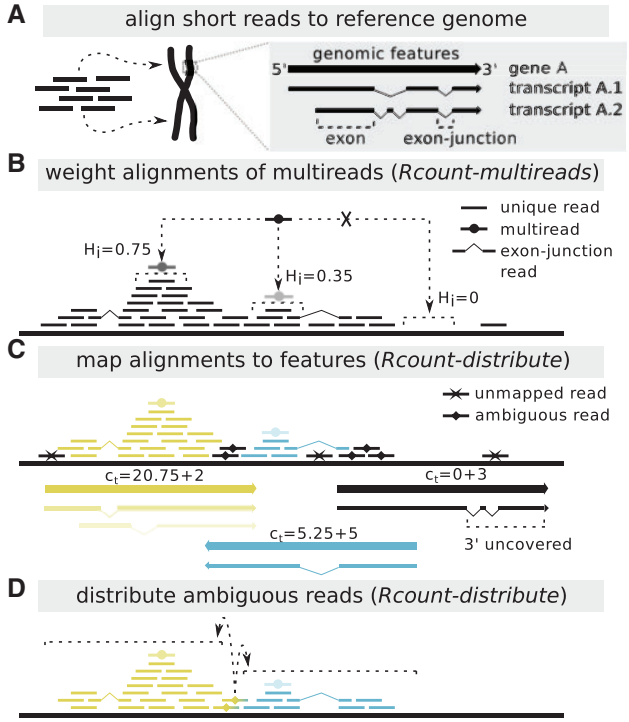


Fig. 1. Schematic Rcount algorithm used to calculate gene expression values. (**A**) After initial quality checks have been performed, the reads are aligned to a reference genome, preferentially with a splice-aware aligner [we tested TopHat2 (Kim *et al.*, 2013), Subread (Liao *et al.*, 2013) and STAR (Dobin *et al.*, 2013)]. (**B**) Alignments of multireads are weighted based on the number of uniquely aligned reads in the neighborhood. (**C**) In a first round, alignments are mapped to all annotated transcripts and treated as unambiguous. Temporary expression values are calculated (c_t) and used to filter the transcripts (optionally, transcripts with uncovered 3' ends are filtered as well). (**D**) In a second round, ambiguous alignments are distributed based on unambiguous alignments, resulting in final expression values (c_f)

During this process, the user can extend the genes (add flanking regions) or remove features from the annotation. Depending on the library preparation protocol, some of the features in the genome annotation are less likely to be sequenced (e.g. rRNA-coding genes with poly(A)-selective library preparation protocols). The user can choose to remove these features or to assign a lower priority to them. If a read aligns to a location where two genes with different priorities overlap, it is automatically assigned to the one with higher priority.

Rcount-distribute sums up the weights of the alignments (hits) per gene in two steps. In the first step, all hits are mapped to all genes (i.e. their transcripts). Transcripts of truly expressed genes should generally have at least some hits in the vicinity of their 3′ end (e.g. due to poly(A)-tail priming during library preparation) and/or at least a minimal total number of hits (user-specified). Transcripts not matching these criteria are discarded during the first round (Fig. 1C). During the second step, the hits are divided into unambiguous and ambiguous. The unambiguous hits are assigned first and subsequently used to proportionally distribute the ambiguous hits (Fig. 1D). The transcripts are re-filtered using the same criteria as before. The final expression value c_f of a gene is then calculated as the sum of hits assigned to any of its transcripts (Fig. 1D).

The final output is one count table per sample. In addition to the final expression values, the output table also contains the number of unambiguous and ambiguous (before and after distributing them) hits per gene (either on the whole gene length, or only within a certain number of bases from the 3′ end of the transcript, which can be specified by the user). To extract a certain column or to merge multiple samples for downstream analyses, an R script is provided on github.com/MWSchmid/Rcount.

ACKNOWLEDGEMENT

We thank Dr. Diana E. Coman Schmid (EAWAG) for helpful discussions and software testing.

Funding: This work was supported by the University of Zurich, and grants from the Swiss National Science Foundation to U.G.

Conflict of interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Anders,S. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protocols*, **8**, 1765–1786.

Chung,D. *et al.* (2011) Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLOS Comput. Biol.*, **7**, e1002111.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Kim,D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

Liao,Y. *et al.* (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.*, **41**, e108.

Liao,Y. *et al.* (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Rapaport,F. *et al.* (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, R95.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 321–332.

Schmid,M.W. *et al.* (2012) A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing. *PLoS One*, **7**, e29685.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

6 HiCdat: a fast and easy-to-use Hi-C data analysis tool

The following manuscript is intended as a software article. I designed the data handling concepts and implemented the pre-processor tool HiCdat-Pre. Stefan Grob and I designed the data analysis concepts underlying HiCdatR together. Stefan drafted some code for HiCdatR. Where necessary, I adapted the code, designed the automization process and implemented the final version of HiCdatR. I wrote the manuscript and the user guide. Stefan Grob critically read the manuscript and the user guide and provided valuable feedback. Ueli Grossniklaus read and corrected the final draft and contributed the final name of the software.

HiCdat: a fast and easy-to-use Hi-C data analysis tool

Marc W Schmid[1,*], Stefan Grob[1], Ueli Grossniklaus[1]
1 Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zürich, Switzerland
*** E-mail: marcschmid@gmx.ch**

Abstract

Background: The study of nuclear architecture using Chromosome Conformation Capture (3C) technologies is a novel frontier in biology. With further reduction in sequencing costs, the potential of Hi-C in describing nuclear architecture as a phenotype is only about to unfold. To use Hi-C for phenotypic comparisons among different cell types, conditions, or genetic backgrounds, Hi-C data processing needs to be more accessible to biologists. **Results:** HiCdat provides a simple graphical user interface for data pre-processing and a collection of higher-level data analysis tools implemented in R. Data pre-processing also supports a wide range of additional data types required for in-depth analysis of the Hi-C data (e.g. RNA-Seq, ChIP-Seq, and BS-Seq). **Conclusions:** HiCdat is easy-to-use and provides solutions starting from aligned reads up to the in-depth analysis. Importantly, HiCdat is focussed on the analysis of larger structural features of chromosomes, their correlation to genomic and epigenomic features, and on comparative studies. It uses simple input and output formats and can therefore easily be integrated into existing workflows or combined with alternative tools.

Background

The development of Chromosome Conformation Capture (3C) techniques and their high throughput derivatives (e.g., 4C and Hi-C) has enabled the analysis of nuclear architecture (i.e. chromatin organization) at an unprecedented resolution [1]. Hi-C data analysis comprises a large variety of approaches, including point-to-point looping interactions (e.g., promoter-enhancer interactions), three-dimensional modeling of chromatin [2], identification of structural domains (e.g., topologically associated domains, TADs [3]), or comparison of different genetic backgrounds (e.g., wild-type *versus* mutant tissues [4–6]).

The large number of reads produced by Hi-C experiments (e.g., around 200-300 mio aligned read-pairs per sample in [3]) requires efficient tools for processing, filtering, and simplification of the data to best match the demands of the downstream analyses. Several open-source tools are available, each with its own scope and requirements. HiCUP [7] performs mapping and quality control on Hi-C data but no downstream analysis. Sushi [8] and HiTC [9] provide data visualization functionality, but no pre-processing or statistical

analysis of Hi-C data. HiCseg specifically focusses on identification of domains in Hi-C data [10]. ChromoR [11] offers data pre-processing and sample comparison, but does not support the analysis of additional genomic and epigenomic features. HiCpipe [12] implements a computationally very intense normalization method, which does not perform better than the parametric approach in HiCNorm [13] (normalization method). HOMER [14] and hiclib [15] offer a large variety of functionalities, including pre-processing and higher-level data analysis. However, these tools may be inaccessible to users with limited programming experience: HOMER requires some command-line skills and only generates plain-text output, which needs to be further processed by the user; hiclib requires familiarity with Python. The latter is less well known among molecular biologists and geneticists who are likely more familiar with R. Alternatively, HiBrowse offers many functionalities in an easy-to-use web-interface [16], which, however, constrains the user by forcing to adhere to the available procedures and the requirement of uploading their data to a web server.

Envisioning nuclear architecture (i.e. chromatin organization) as an ordinary phenotype of an organism or a specific tissue type (e.g. like the transcriptome), comparative Hi-C experiments may soon be of very broad interest, raising the need for data analysis tools that are not only well-accessible to bioinformaticians. We therefore developed HiCdat. It includes a fast and easy-to-use GUI tool for Hi-C data pre-processing and an R [17] package, which implements all data analysis approaches employed in [5].

Implementation

HiCdat was developed with a focus on speed, user-friendliness, and flexibility in terms of file formats. The GUI tool for data pre-processing serves to convert large-scale genomic and epigenomic data into simple tables, which can be efficiently loaded and processed within R. The R-package provides a collection of functions, which allow higher-level data analysis (e.g., as in [5]) with only few lines of codes. Data formats are kept as simple as possible to ensure that the user can easily integrate HiCdat into a pre-existing workflow or combine it with other tools.

Results and Discussion

HiCdat is divided into two parts (Figure 1): (i) a GUI tool for data pre-processing (termed *HiCdatPre*) and (ii) an R-package for higher-level data analysis (termed *HiCdatR*).

Data pre-processing with *HiCdat*

HiCdat takes as input two alignment files (forward and reverse reads, hereafter termed read-ends) in BAM format (Binary Alignment/Map), a reference genome, and various data types from additional experiments (e.g., genome annotation, RNA-Seq, ChIP-Seq, BS-Seq data). There are five automated steps during data pre-processing: (i) merging of reads, (ii) creating fragments, (iii) mapping of read-ends to fragments, (iv) processing data from additional experiments, and (v) creating organism-specific R-code.

Merging reads

The read-ends are first aligned seperately to the reference genome using, for example, Subread [18]. Uniquely aligning read-ends are then merged based on their common read name (around 12.6 million read-ends per minute[a]).

Creating fragments

Hi-C data analysis can either be carried out on restriction fragments or genomic bins with fixed size. Both types of fragments can be created by supplying the reference genome sequence and one or more restriction enzymes or a fixed bin size.

Mapping read-ends to fragments

To calculate the interaction frequency between two fragments, the merged read-pairs are first mapped to the fragments' coordinates and then summarized as number of interactions per fragment pair (around 7.5 million read-pairs per minute[a]). During this procedure, the read-pairs can optionally be filtered using the approach proposed by [19]. Read-pairs with each end aligning at the opposite strand are thereby removed if they are too close to each other. There are two cases: (i) A read-pair where the two ends point towards each other ("inward-pair"), and (ii) a read-pair where the two ends point away from each other ("outward-pair"). Inward-pairs spanning only a short region may be caused by uncut DNA. Outward-pairs spanning only a short region can be a result of self-ligation.

Processing data from additional experiments

To analyze the interplay between the Hi-C interactome and genomic/epigenomic features, a large variety of such information can be automatically added to the fragments. In principle there are two fundamentally different types of data: counts and densities. During higher-level data analysis, counts are generally log-transformed, whereas densities are kept as percentages. Likewise, if data are summarized over multiple fragments (e.g. to obtain the annotation for 1 Mb bins directly from the fragment annotation), counts are summed up, whereas densities are averaged. Both data types comprise two sub-types, resulting in four different types of "tracks" which can be processed: (i) genome annotation features

(e.g., genes and transposons), (ii) short count features (e.g., RNA-Seq and smallRNA-Seq), (iii) density features (e.g., ChIP-Seq), and (iv) DNA-methylation density (e.g., BS-Seq).

Genome annotation features (GFF/GTF files with multiple feature types per file) can generally be very long and possibly span multiple fragments. The number of elements per fragment is therefore counted as follows: If the feature spans the entire fragment, a value of 1 is added. If the feature only partly overlaps (or is within) the fragment, a value of 0.5 is added. In contrast, short count features (BAM files with one feature type only) are mostly entirely within a fragment and are therefore simply summed up per fragment.

Density of a certain feature (BAM files with one feature type only) is calculated as the number of bases covered by at least one element (e.g. short read) divided by the length of the fragment (times 100 to obtain percentages). Likewise, DNA cytosin-methylation density corresponds to the percentage of methylated C's per fragment.

Creating the organism-specific R-code

The higher-level data analysis requires some organism-specific R-code, which can be obtained by supplying the reference genome sequence and the restriction enzyme(s) used for the Hi-C library preparation.

Data analysis with *HiCdatR*

In-depth Hi-C data analysis is done in R with *HiCdatR*. The only inputs required are the interaction counts per fragment pair and, optionally, the annotation of the fragments holding the genomic and epigenomic tracks. For most of the functions, it is furthermore possible to supply tables specifying genomic regions of interest (e.g. chromosome arms or pericentromeres as in [5]). The functionalities include (i) data normalizations as proposed by [13, 20, 21], (ii) sample correlation matrices, (iii) data visualization, (iv) sample comparisons, (v) calculation of distance decay exponents, (vi) principle component analysis (PCA) including correlation of the first principle component to genomic and epigenomic features, (vii) test for increased interaction frequencies between genomic regions of interest compared to randomly sampled regions, and (viii) test for enrichment or depletion of genomic and epigenomic features within genomic regions of interest compared to randomly chosen regions.

Data normalization

Multiple data normalization strategies have been proposed and implemented in various languages and packages [11–13, 15, 20, 21]. Three of them have been re-implemented in

HiCdat: (i) the distance (*intra*-chromosomal interactions) and coverage (*inter*-chromosomal interactions) normalization described in [20], (ii) the iterative coverage normalization proposed by [21], and (iii) the more sophisticated but highly efficient, normalization using Poisson regression as implemented in HiCNorm [13], which performs similar or better [11, 13] than the procedures from [12, 15].

Sample correlation

To visualize the similarities between samples and replicates, HiCdat uses sample correlation matrices. Correlation between two samples is thereby calculated as the average, or median, correlation between all the individual bins of the interaction matrices (i.e. the virtual 4C tracks, see Additional file 1: Figure S1).

Data visualization

Hi-C interaction frequencies and differences between multiple samples are visualized as heatmap-like images. Individual samples can either be displayed natively (i.e. with their normalized interaction frequencies, Additional file 2: Figure S2) or in a correlated manner (Additional file 3: Figure S3).

Sample comparison

Three different approaches to compare two samples to each other are implemented. In a first approach, the difference of a given fragment pair between the two samples is divided by the average interaction frequency among the two samples resulting in "relative differences" [4] (Additional file 4: Figure S4). Considering that neighboring genomic regions are physically linked to each other, it is likely that they change accordingly. To visualize these domains, the relative differences can be correlated to each other ("correlated differences", Additional file 5: Figure S5). The disadvantage of these approaches is that they rely on visual inspection of the difference matrices. To estimate the significance of the difference and identify the affected regions, we introduced signed difference matrices (SDMs) [5]. Additionally, they also provide an overall estimate of the extent and significance of the difference between two samples (Additional file 6: Figure S6).

Calculation of distance decay exponents

The extent to which interaction frequencies change dependent on the distance to a given point in the genome can be characterized with the interaction decay exponent (IDE). IDEs are calculated as the slope of a linear fit to the average interaction frequencies observed at given distances (both log-transformed, Additional file 7: Figure S7). IDEs were initially used to predict the folding principles of the human genome using two polymer-folding models (the fractal and equilibrium globule module, respectively), which result in

distinct values for the expected IDE [20]. Alternatively, they can also be used to describe differences between certain sub-compartments of the genome, or between samples [5].

Identification of structural domains using principle component analysis (PCA)

The correlation between the interactomes of different genomic regions can be used to identify larger compartments [20] or structural domains [5]. The approach relies on principal component analysis (PCA) of the distance-normalized and correlated *intra*-chromosomal interactions (Additional file 8: Figure S8). The first principal component (FPC) can then be used to differentiate for example the A and B compartments in *Homo sapiens* [20], or loose and compact structural domains in *Arabidopsis thaliana* [5]. The interplay between the FPC and the epigenomic/genomic landscape can be analyzed with two methods: (i) either by using the built-in cor.test() [17] function to test for significance of correlation between FPC and the density/count of a given feature (Additional file 9: Figure S9), or (ii) by using an approach in which the fragments are split into two groups according to the sign of the FPC (Additional file 10: Figure S10, Additional file 11: Figure S11). Enrichment of a given feature can then be calculated as the ratio of the average density/count in one over the other group, and tested for significance using a two-sided Wilcoxon rank sum test [5].

Testing selected regions for increased interaction frequency and enrichment/depletion of epigenomic/genomic features

Given a set of genomic regions of interest, HiCdat can test for increased interaction frequencies between the regions of interest compared to randomly sampled regions. Considering that the interactome can be strongly influenced by the linear position of a certain region along the chromosome (e.g. close to telomere or centromere), or the chromosome number itself [5, 22], random sampling is performed in a "balanced" fashion: Within each random set, the randomly chosen regions reflect the numbers, as well as the locations, of the regions of interest. The procedure creates an empirical distribution of interaction frequencies between random sets, which can then be used to calculate an empirical P-value (one-sided) for the enrichment of interactions between the sets of interest [5]. The same sampling approach can be applied to test for enrichment or depletion of epigenomic or genomic features within a set of genomic regions of interest.

Conclusions

In short, HiCdat allows rapid Hi-C data analysis as described in [5], requiring only little programming experience. The focus lies on the identification of larger structural features of chromosomes, their interplay with the epigenomic/genomic landscape, and on comparative studies. Input and output is kept as simple as possible to enable easy integration

into pre-existing workflows, or the combination of a part of the tool with another tool.

Availability and requirements

Project name: HiCdat

Project home page: github.com/MWSchmid/HiCdat

Operating systems: Windows (7), MacOSX (10.9), Ubuntu-like Linux distributions (all 64 bit)

Programming language: C++ and R

Other requirements: R-packages: randomizeBE and gplots

License: GNU GPL v3

Any restrictions to use by non-academics: None

List of abbreviations

GFF: General Feature Format, GTF: Gene Transfer Format, BAM: Binary Alignment Map, IDE: Interaction Decay Exponent, PCA: Principle Component Analysis, FPA: First Principle Component.

Run-times and test-system

[a] Run-times were measured on a 64 bit Kubuntu running on an Intel Core i7 930@2.8 GHz with 24 Gb RAM and a 7'200 rpm Samsung HDD using Hi-C data from mouse embryonic stem cell (GSM862720, GSM862721) and cortex (GSM862750, GSM862751) samples from [3] (NCBI37 assembly, and 1 Mb bins for the higher-level data analysis, and 823'377 *Hind*III restriction fragments for mapping to fragments).

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Designed and implemented the pre-processing tool: MWS. Designed and implemented higher-level data analysis in R: MWS SG. Wrote the manuscript: MWS. Helped to write manuscript: SG UG.

Acknowledgements

This work was supported by the University of Zurich, an iPhD project of SystemsX.ch, the Swiss Initiative in Systems Biology, and grants from the Swiss National Science Foundation (SNF) and the European Research Council (ERC) to UG.

References

[1] Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14: 390-403.

[2] Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, et al. (2010) A three-dimensional model of the yeast genome. Nature 465: 363-367.

[3] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485: 376-380.

[4] Moissiard G, Cokus SJ, Cary J, Feng S, Billi AC, et al. (2012) MORC family ATPases required for heterochromatin condensation and gene silencing. Science 336: 1448-1451.

[5] Grob S, Schmid MW, Grossniklaus U (2014) Hi-C analysis in *Arabidopsis* identifies the *KNOT*, a structure with similarities to the *flamenco* locus of *Drosophila*. Mol Cell 55: 678-693.

[6] Feng S, Cokus SJ, Schubert V, Zhai J, Pellegrini M, et al. (2014) Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in *Arabidopsis*. Mol Cell 55: 694-707.

[7] Hicup. URL `http://www.bioinformatics.babraham.ac.uk/projects/hicup/`.

[8] Phanstiel DH, Boyle AP, Araya CL, Snyder MP (2014) Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. Bioinformatics 30: 2808-2810.

[9] Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen C, et al. (2012) HiTC: exploration of high-throughput 'C' experiments. Bioinformatics 28: 2843-2844.

[10] Hicseq r-package. URL `http://cran.r-project.org/web/packages/HiCseg/index.html`.

[11] Shavit Y, Lio' P (2014) Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. Mol Biosyst 10: 1576-1585.

[12] Yaffe E, Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet 43: 1059-1065.

[13] Hu M, Deng K, Selvaraj S, Qin ZS, Ren B, et al. (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics 28: 3131-3133.

[14] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38: 576-589.

[15] Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, et al. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods 9: 999-1003.

[16] Paulsen J, Sandve GK, Gundersen S, Lien TG, Trengereid K, et al. (2014) HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. Bioinformatics 30: 1620-1622.

[17] R-project. URL `http://www.r-project.org/`.

[18] Liao Y, Smyth GK, Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 41: e108.

[19] Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, et al. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503: 290-294.

[20] Liebermann-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289-293.

[21] Zhang Y, McCord RP, Ho YJ, Lajoie BR, Hildebrand DG, et al. (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell 148: 908-921.

[22] Grob S, Schmid MW, Grossniklaus U (2013) Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture. Genome Biol 14: R129.

[23] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

[24] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.

[25] Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, et al. (2007) Genome-wide profiling and analysis of *Arabidopsis* siRNAs. PLoS Biol 5: e57.

[26] Gregory BD, O'Malley RC, Lister R, Urich MA, Tonti-Filippini J, et al. (2008) A link between RNA metabolism and silencing affecting *Arabidopsis* development. Dev Cell 14: 854-866.

[27] Lister R, O'Malley RC, Tonti-Filippini J, Gregory BG, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133: 523-536.

[28] Jacob Y, Stroud H, LeBlanc C, Feng S, L Z, et al. (2010) Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. Nature 466: 987-991.

[29] Luo C, Sidote DJ, Zhang Y, Kerstetter RA, Michael TP, et al. (2013) Integrative analysis of chromatin states in *Arabidopsis* identified potential regulatory mechanisms for natural antisense transcript production. Plant J 73: 77-90.

[30] Stroud H, Greenberg MVC, Feng S, Bernatavichute YV, Jacobsen SE (2013) Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. Cell 152: 352-364.

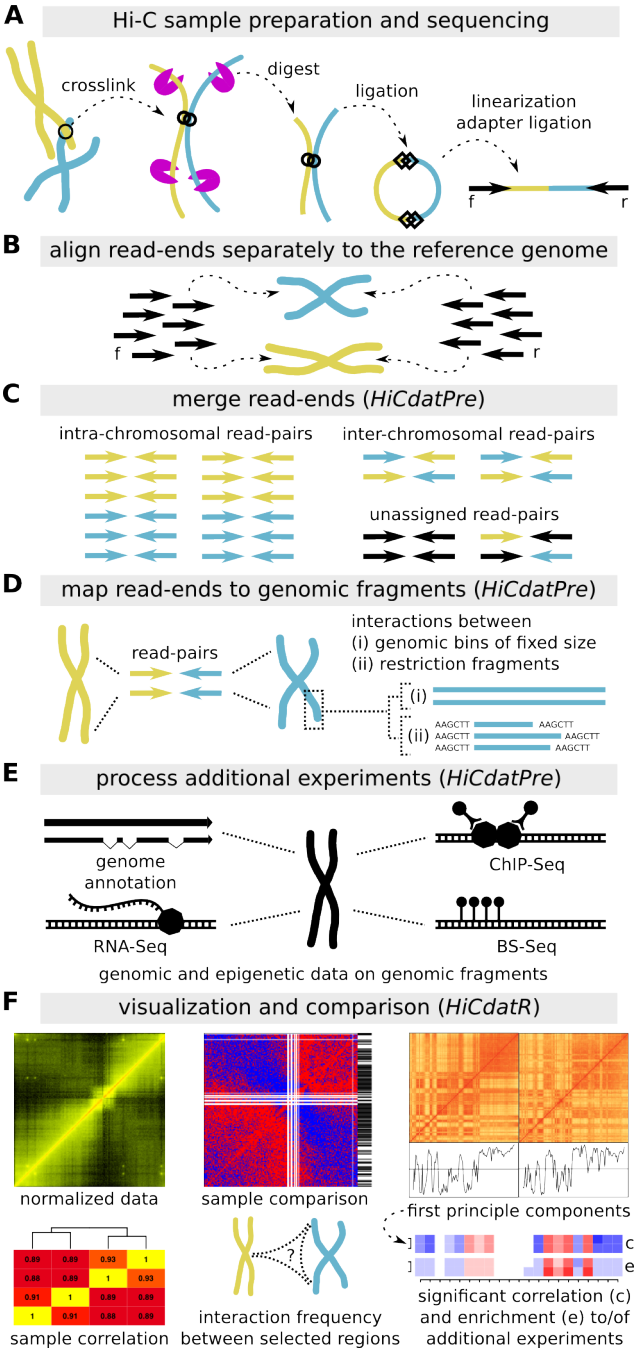


Figure 1. Schematic HiCdat workflow. (**A-B**) After sequencing and initial quality checks have been performed, the read-ends (f: forward, r: reverse) are aligned separately to a reference genome. (**C-D**) After merging the separated read-ends, each end is mapped to genomic fragments, which are either genomic bins with a fixed size or restriction fragments with variable size. (**E**) Genomic fragments can be associated with various data types to test for correlation and enrichment of Hi-C data with genomic and epigenomic features. (**F**) Finally, the data can be conveniently analyzed in R using HiCdatR.

Additional Files

Additional file 1 — Figure S1

Correlation between five samples of *Arabidopsis thaliana* seedlings [4,5] aligned with either Bowtie [23], Bowtie 2 [24], or Subread [18], and processed with either HiCdat or hiclib [15] using a resolution of 100 kb.

Additional file 2 — Figure S2

Visualization of Hi-C interaction frequencies in a pooled wild-type sample of *A. thaliana* [4, 5] (100 kb bins).

Additional file 3 — Figure S3

Visualization of distance-normalized and correlated Hi-C interaction frequencies in a pooled wild-type sample of *A. thaliana* [4, 5] (100 kb bins).

Additional file 4 — Figure S4

Enrichment (blue) and depletion (red) of interaction frequencies in the wild-type compared to the *crowded nuclei4* (*crwn4*) mutant sample of *A. thaliana* [5] (100 kb bins).

Additional file 5 — Figure S5

Correlation of differences between the wild-type and the *crwn4* mutant samples of *A. thaliana* [5] (100 kb bins).

Additional file 6 — Figure S6

Visualization of the difference between the wild-type and *crwn4* mutant samples of *A. thaliana*, [5] using the signed difference matrix (100 kb bins).

Additional file 7 — Figure S7

Distance-dependent decay of interaction frequencies along entire chromosomes in a pooled wild-type sample of *A. thaliana* [4, 5] (100 kb bins).

Additional file 8 — Figure S8

Visualization of distance-normalized and correlated Hi-C interaction frequencies (top), the resulting first principle component (mid), and the distribution of the correlation values (bottom). Data shown for the right arms of chromosomes 1, 4, and 5 from a pooled wild-type sample of *A. thaliana* [4, 5] (100 kb bins).

Additional file 9 — Figure S9

Significant correlation (blue: positive, red: negative) of the first principle component with various genomic and epigenomic features. Data shown for the right arms of chromosomes 1, 4, and 5 from a pooled wild-type sample of *A. thaliana* [4, 5] (100 kb bins). Additional data from www.arabidopsis.org and [25–30].

Additional file 10 — Figure S10

Significant enrichment (blue) and depletion (red) of genomic and epigenomic features in regions with positive Eigenvalues compared to regions with negative Eigenvalues. Data shown for the right arms of chromosomes 1, 4, and 5 from a pooled wild-type sample of *A. thaliana* [4, 5] (100 kb bins). Additional data from www.arabidopsis.org and [25–30].

Additional file 11 — Figure S11

Distribution of epigenomic and genomic features in the structural domains with either positive (blue) or negative (red) Eigenvalues. Data from www.arabidopsis.org and [25–30].

Sample correlation

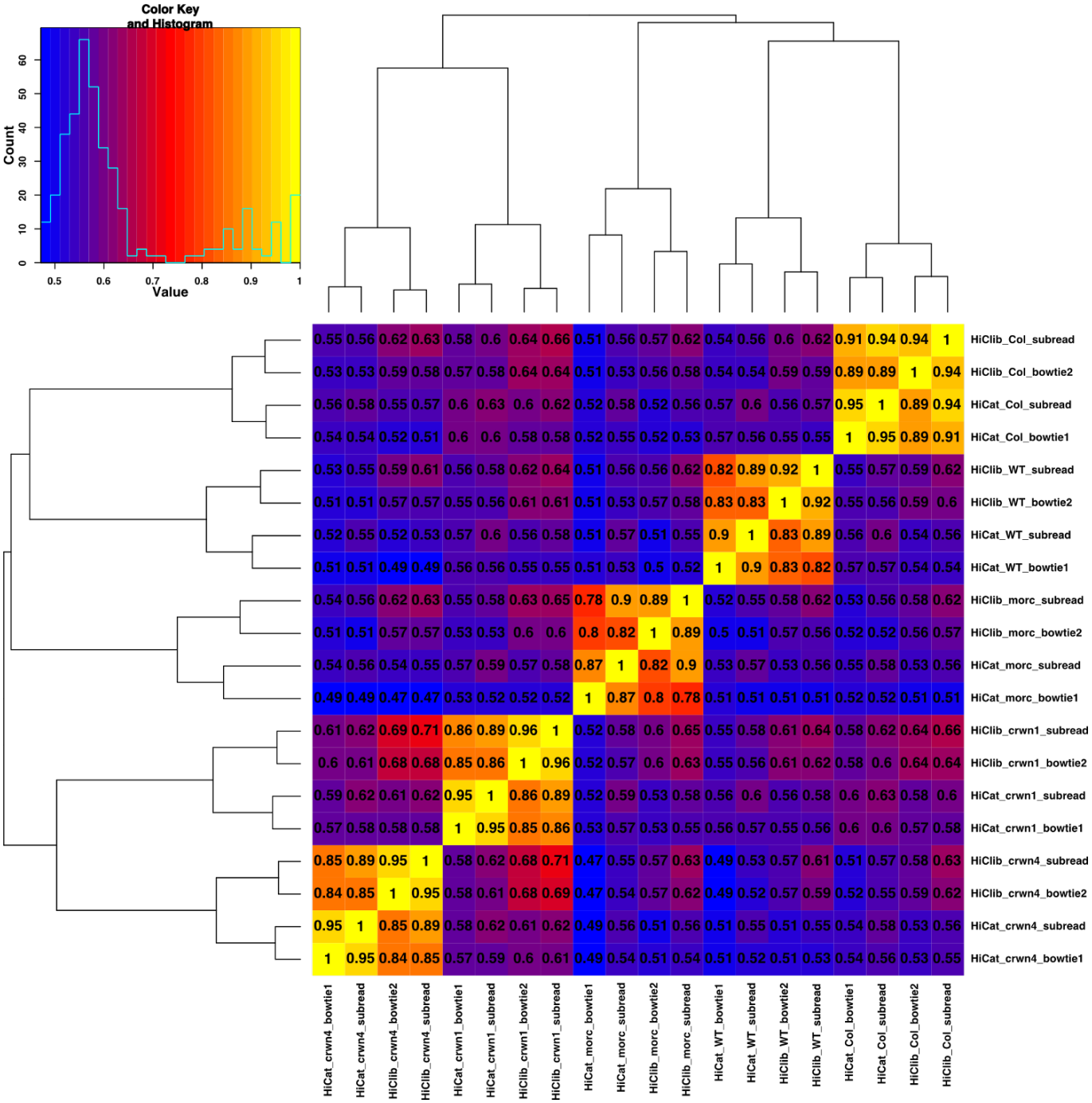


Figure S1: Correlation between five samples of *Arabidopsis thaliana* seedlings [1, 2] aligned with either bowtie [3], bowtie 2 [4], or Subread [5] and processed with either HiCat or hiclib [6] using a resolution of 100 kb.

Data visualization

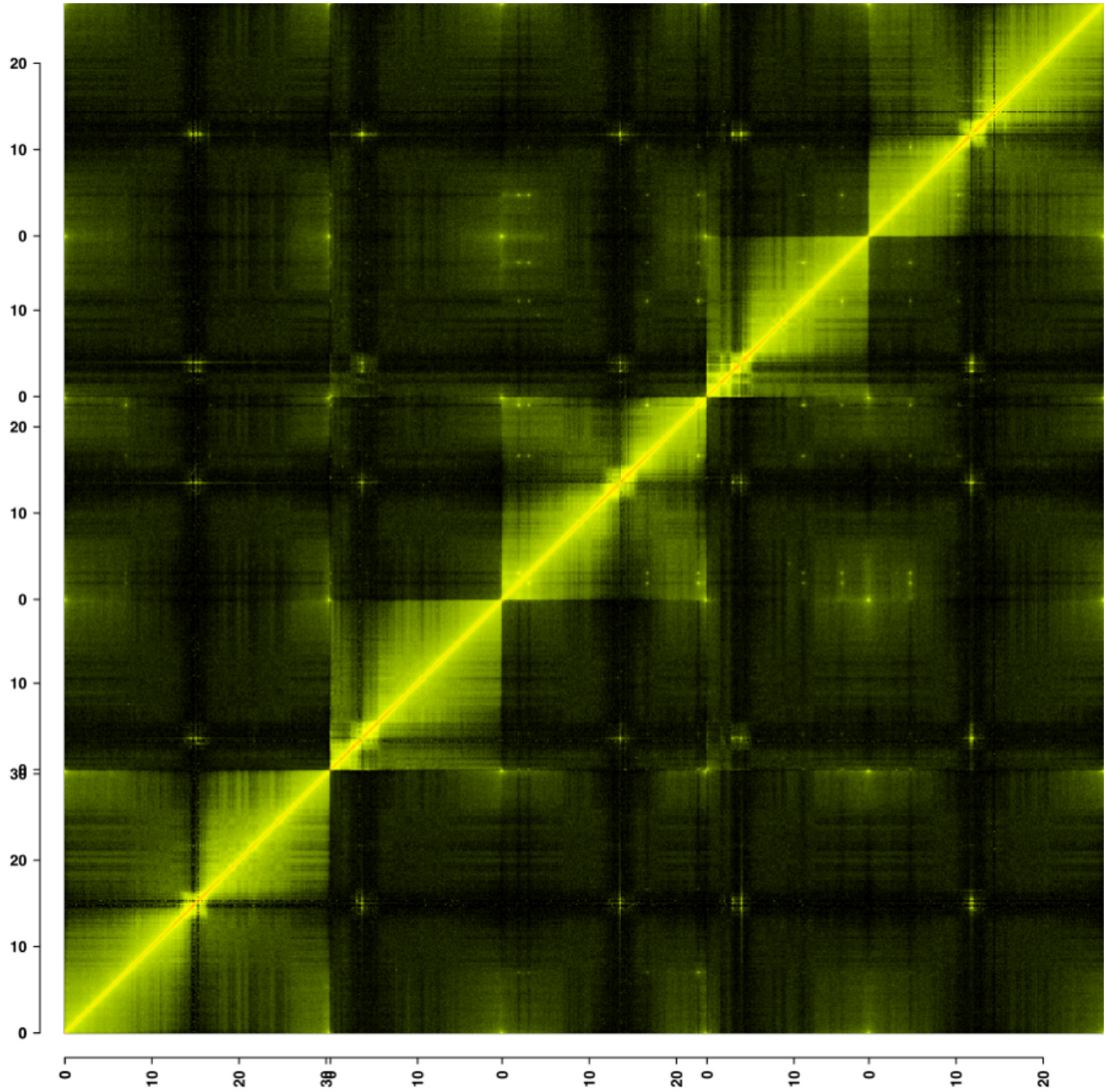


Figure S2: Visualization of Hi-C interaction frequencies in a pooled wild-type sample of *A. thaliana* [1, 2] (100 kb bins).

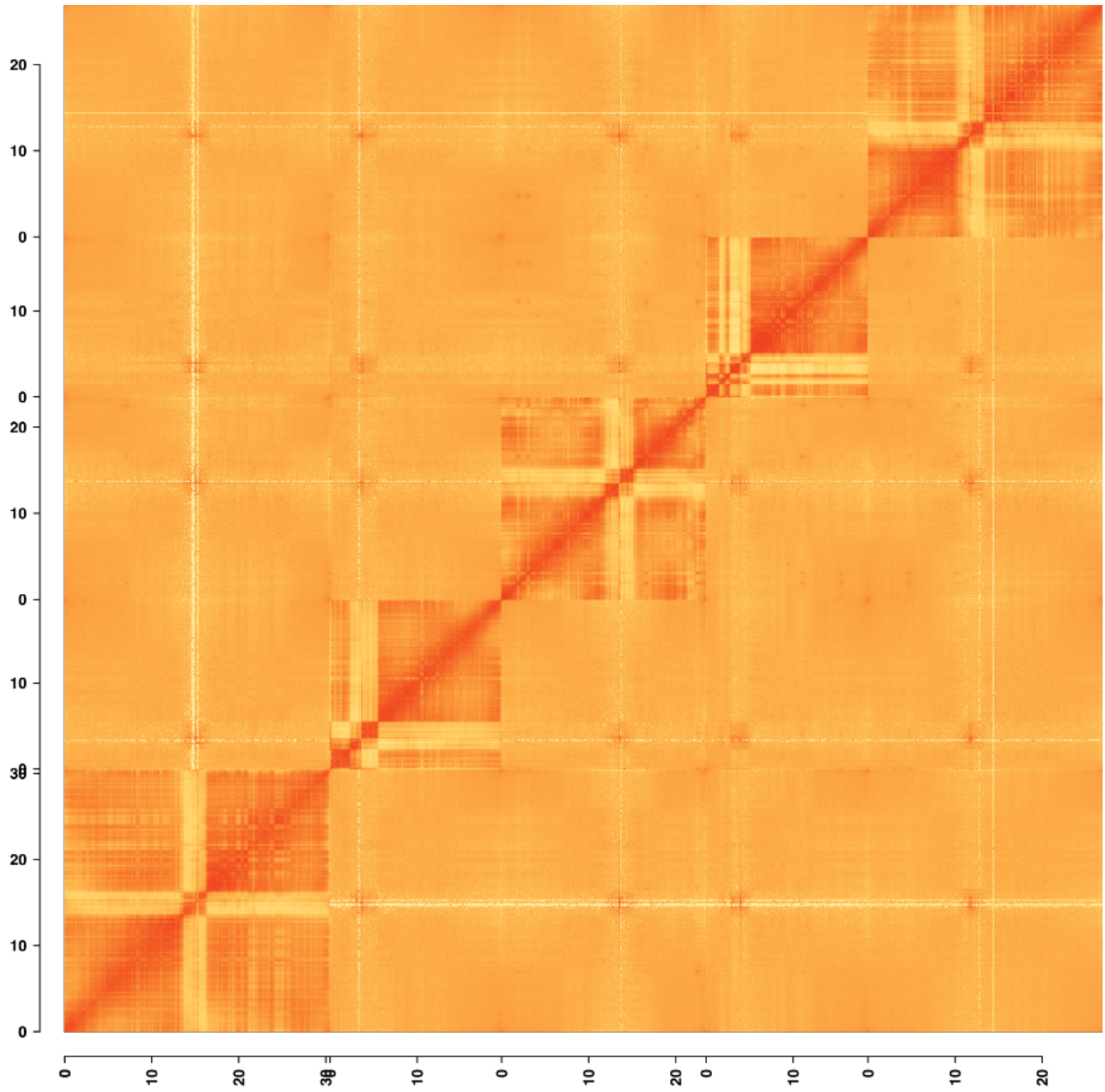


Figure S3: Visualization of distance-normalized and correlated Hi-C interaction frequencies in a pooled wild-type sample of *A. thaliana* [1, 2] (100 kb bins).

Sample comparison

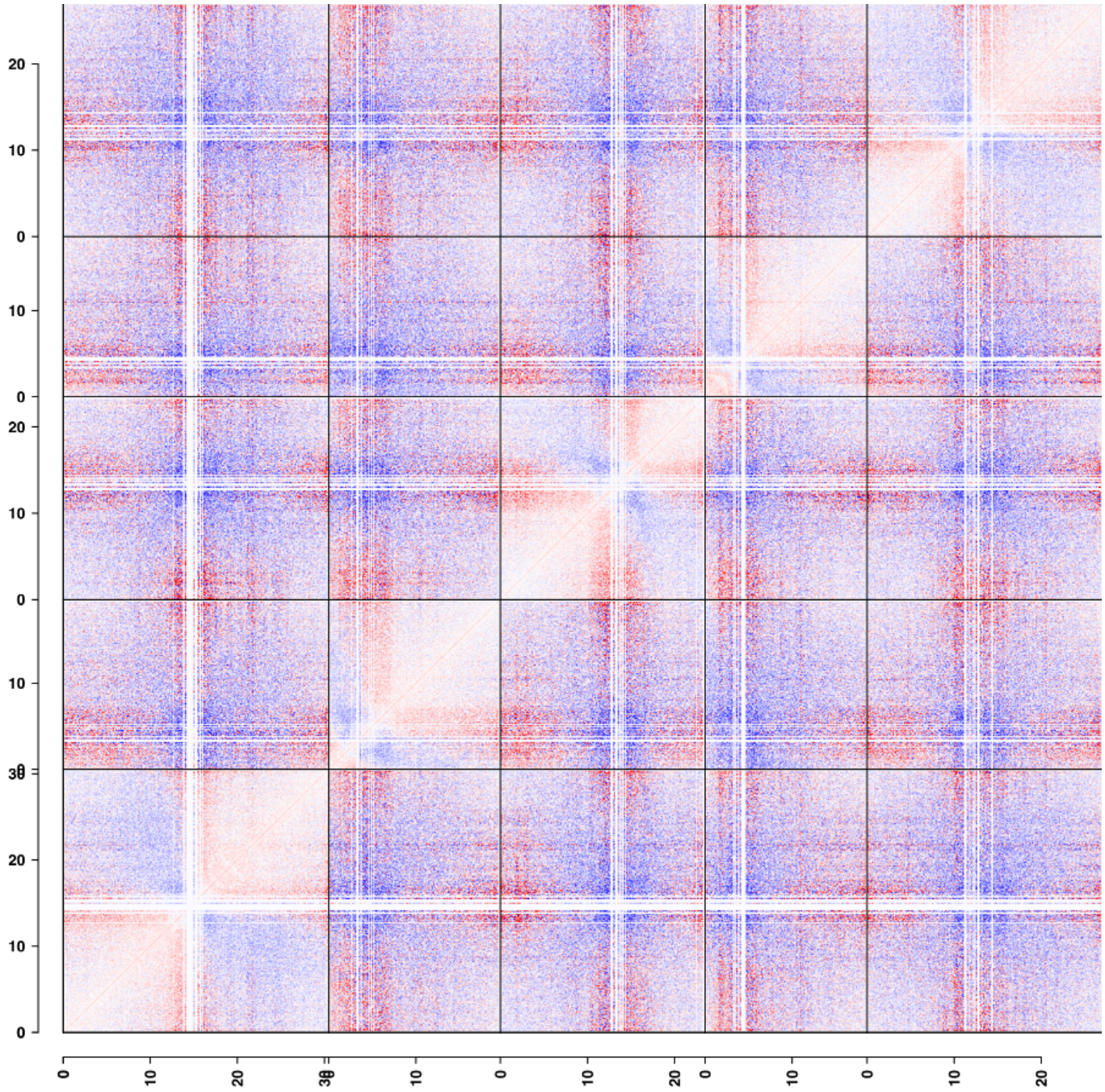


Figure S4: Enrichment (blue) and depletion (red) of interaction frequencies in the wild-type compared to the *crwn4* mutant sample of *A. thaliana* [2] (100 kb bins).

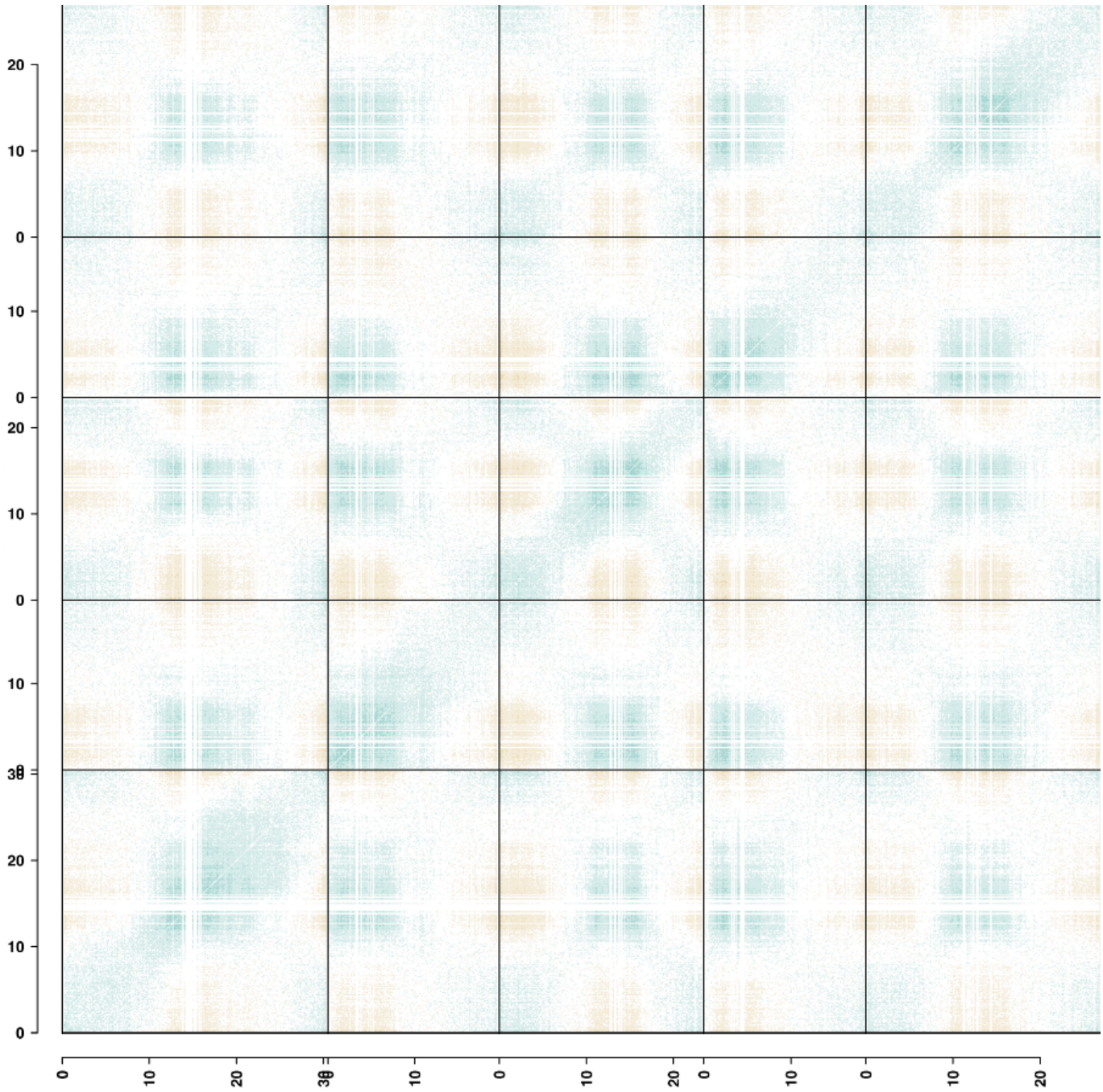


Figure S5: Correlation of differences between the wild-type and the *crwn4* mutant samples of *A. thaliana* [2] (100 kb bins).

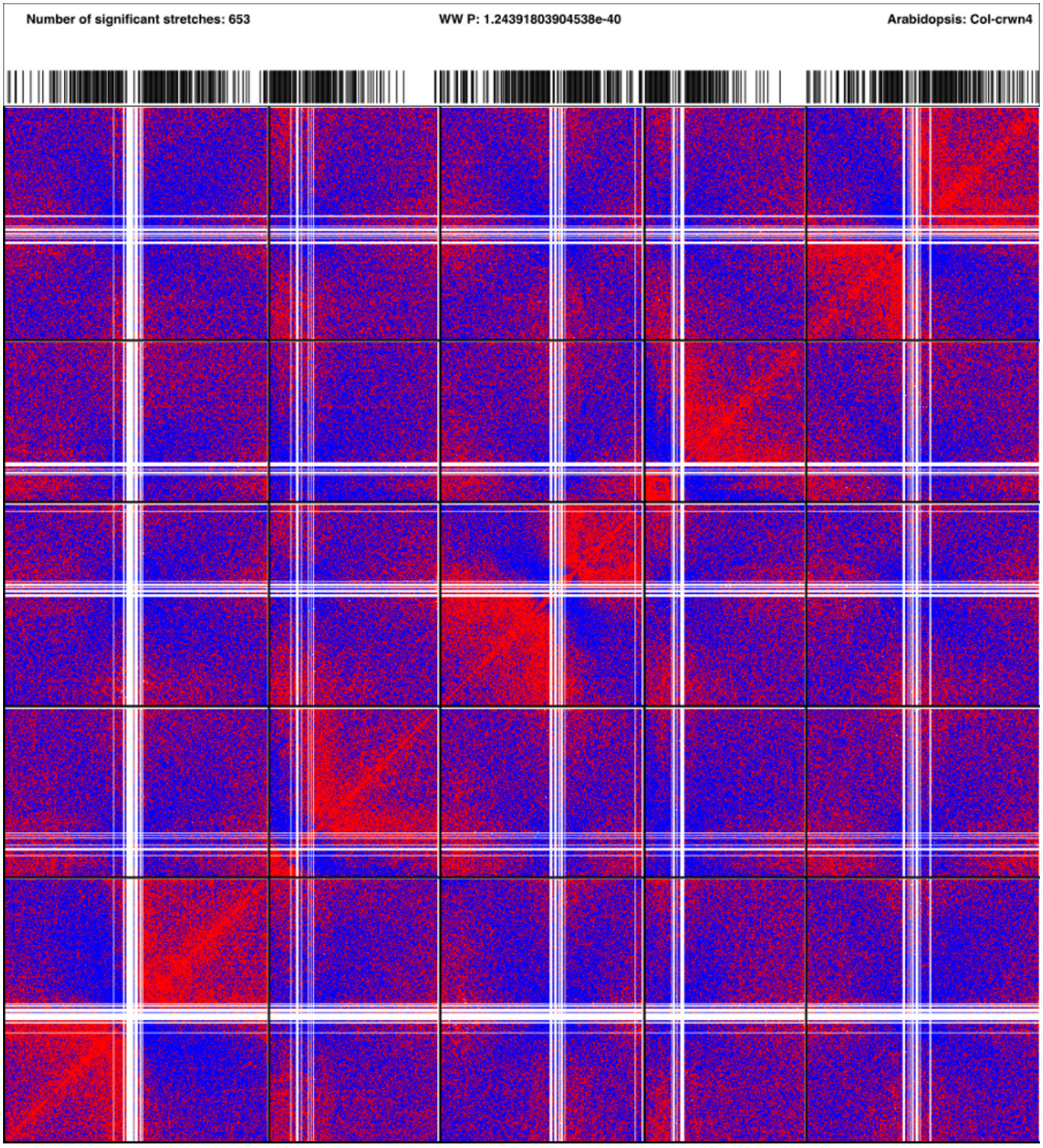


Figure S6: Visualization of the difference between the wild-type and *crwn4* mutant samples of *A. thaliana* [2] using the signed difference matrix (100 kb bins).

Distance decay exponent

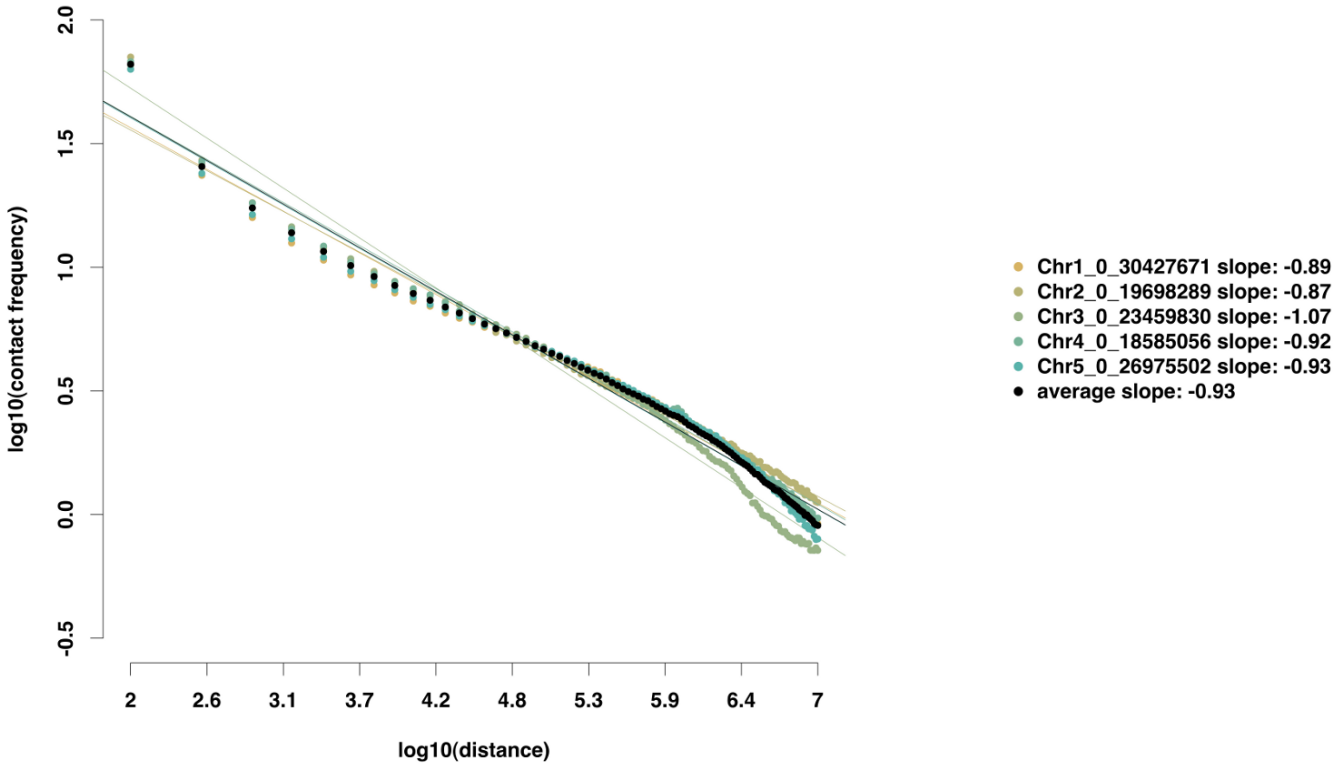


Figure S7: Distance-dependent decay of interaction frequencies along entire chromosomes in a pooled wild-type sample of *A. thaliana* [1, 2] (100 kb bins).

Principle component analysis (PCA)

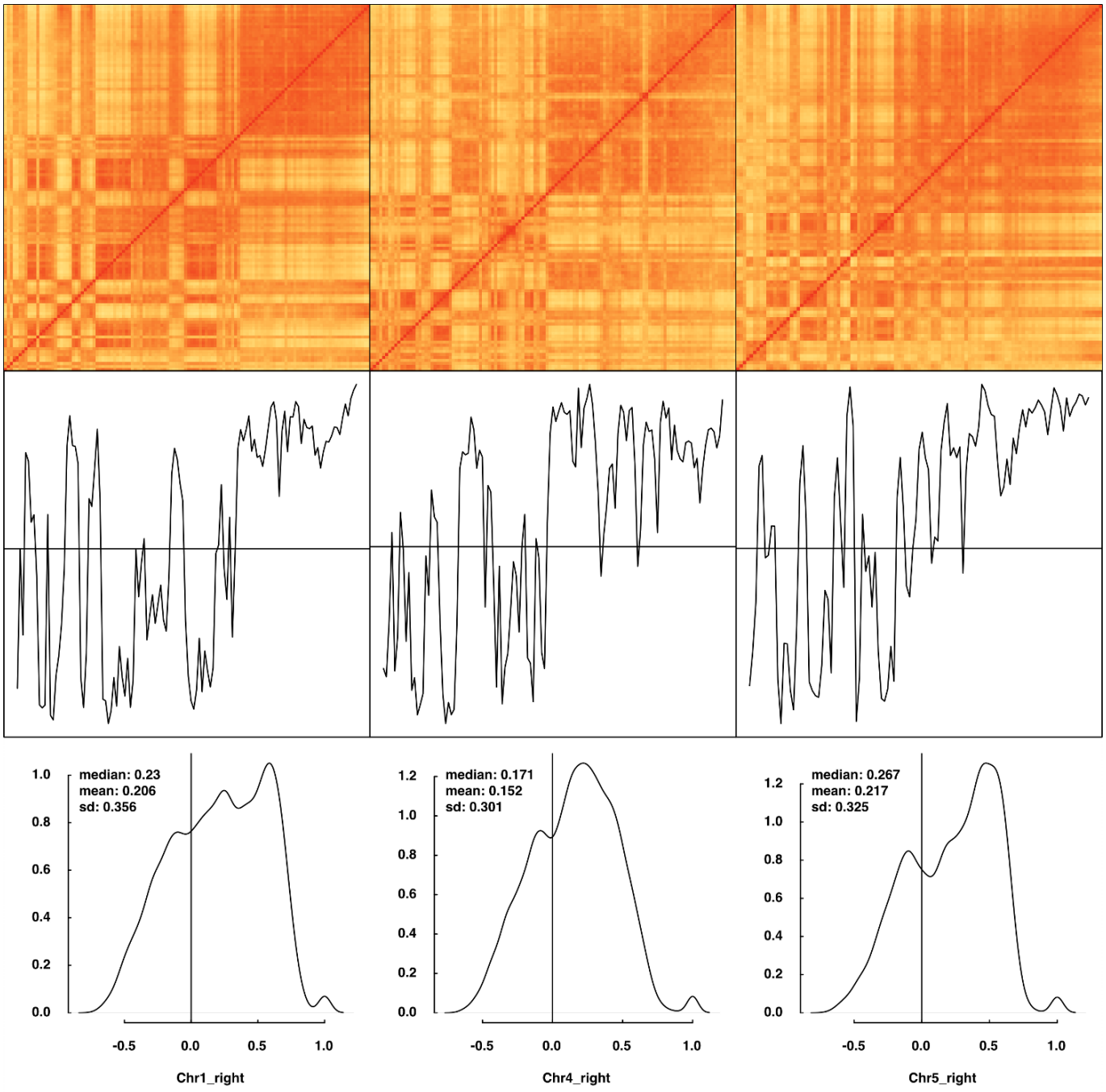


Figure S8: Visualization of distance-normalized and correlated Hi-C interaction frequencies (top), the resulting first principle component (mid), and the distribution of the correlation values (bottom). Data shown for the right arms of chromosomes 1, 4, and 5 from a pooled wild-type sample of *A. thaliana* [1, 2] (100 kb bins).

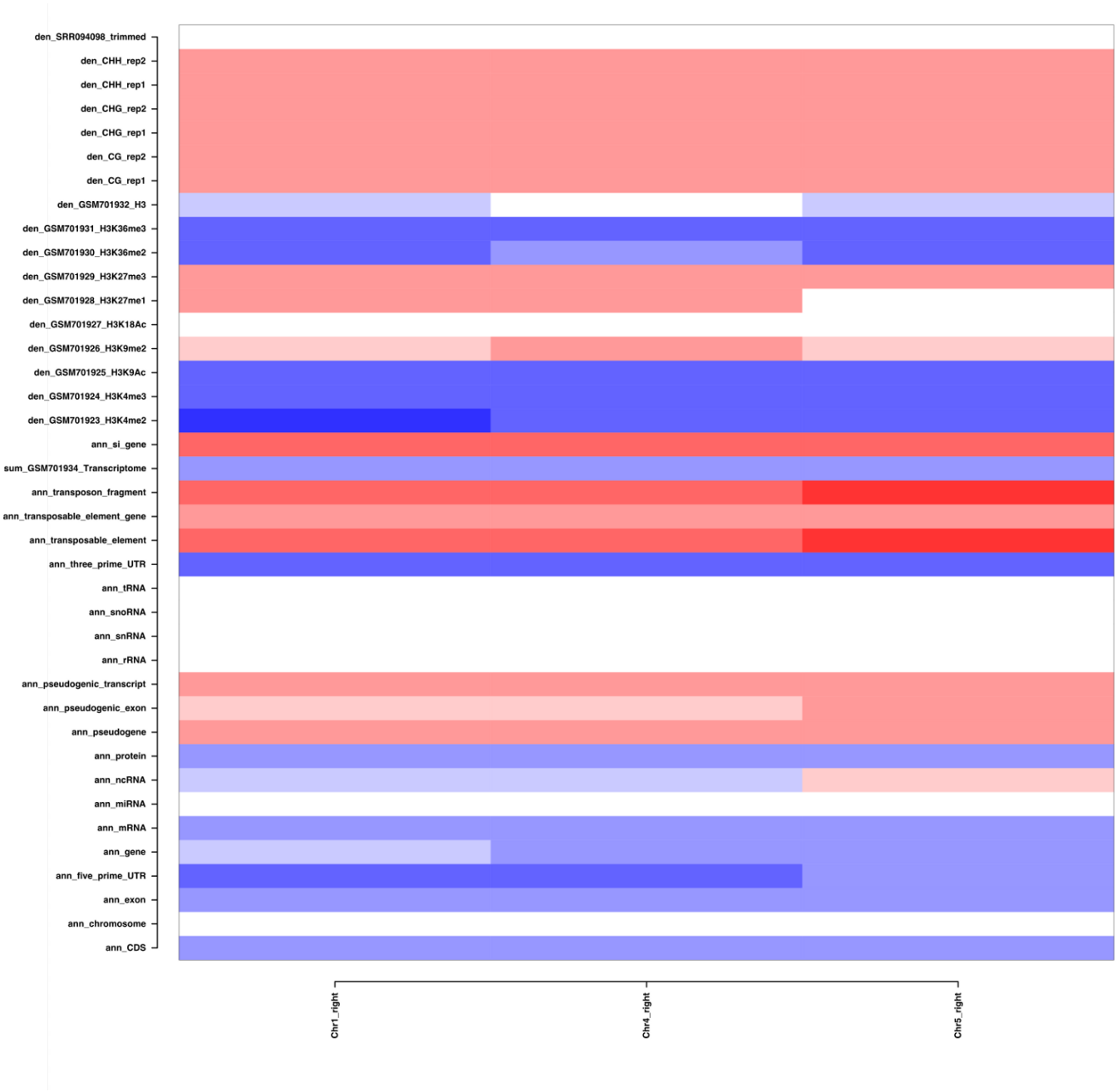


Figure S9: Significant correlation (blue: positive, red: negative) of the first principle component with various genomic and epigenetic features. Data shown for the right arms of chromosomes 1, 4, and 5 from a pooled wild-type sample of *A. thaliana* [1, 2] (100 kb bins). Additional data from www.arabidopsis.org and [7–12].

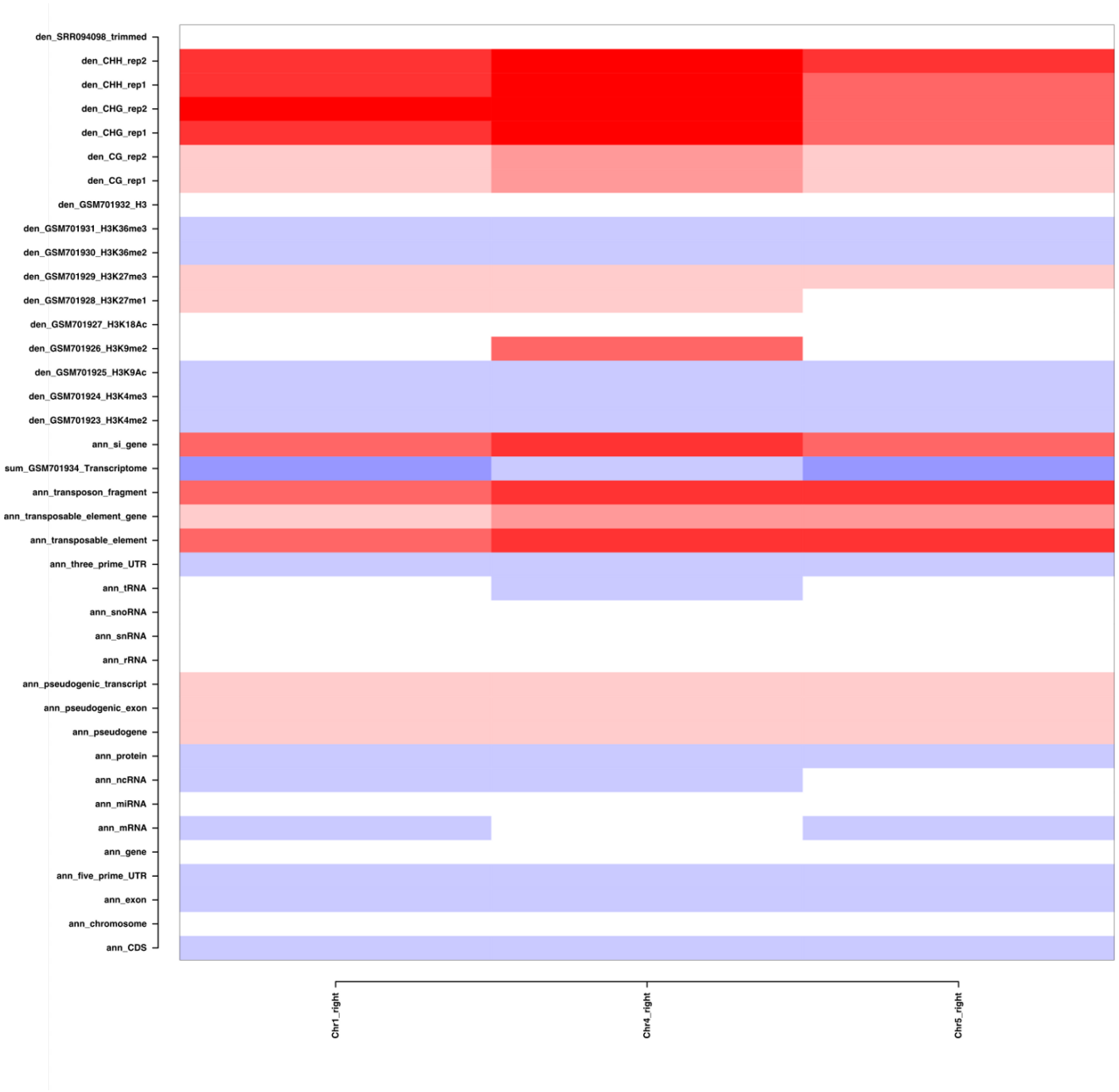


Figure S10: Significant enrichment (blue) and depletion (red) of genomic and epigenetic features in regions with positive Eigenvalues compared to regions with negative Eigenvalues. Data shown for the right arms of chromosomes 1, 4, and 5 from a pooled wild-type sample of *A. thaliana* [1,2] (100 kb bins). Additional data from www.arabidopsis.org and [7–12].

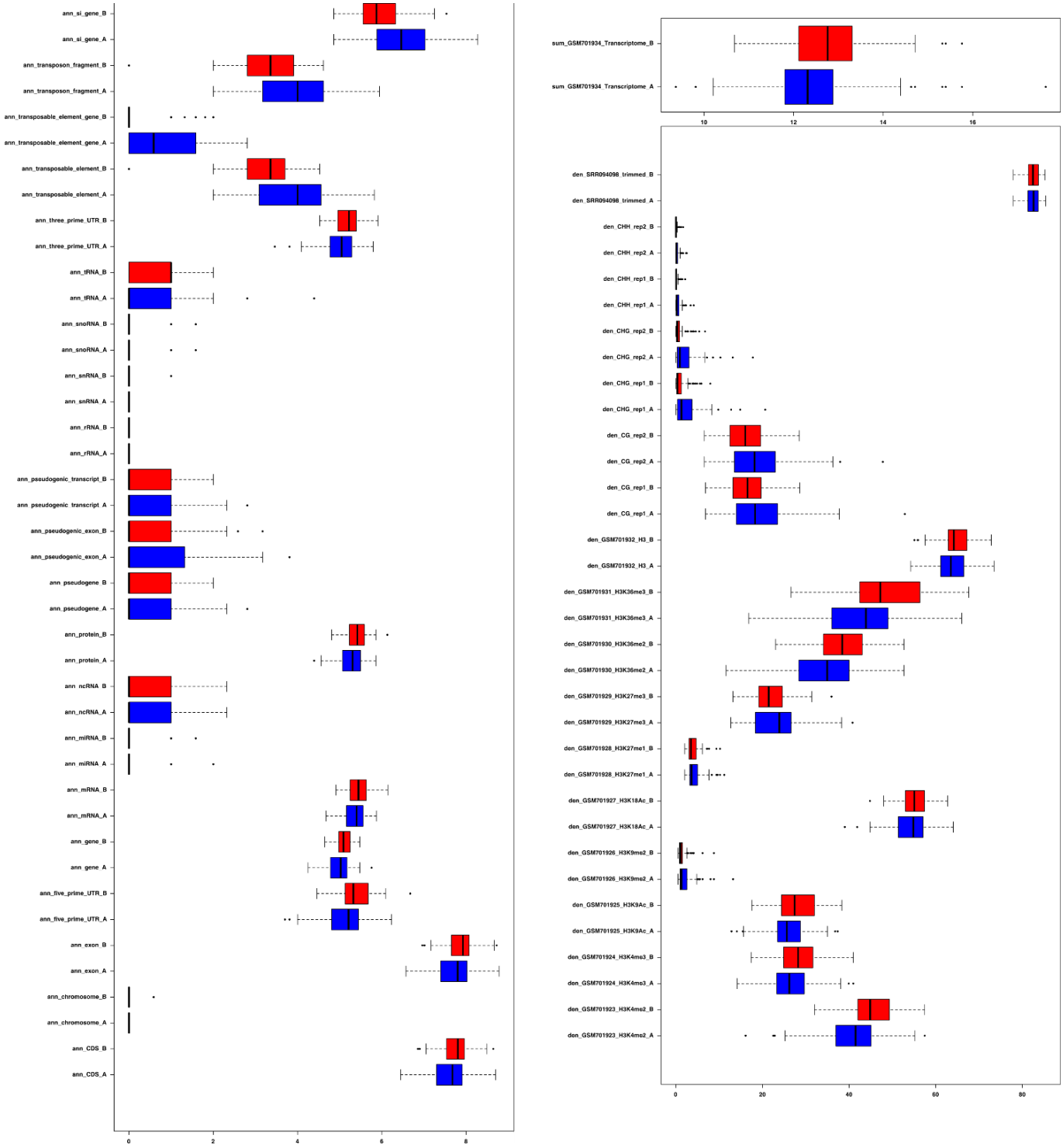


Figure S11: Distribution of epigenetic and genomic features in the structural domains with either positive (blue) or negative (red) Eigenvalues. Data from www.arabidopsis.org and [7–12].

References

[1] Moissiard G, Cokus SJ, Cary J, Feng S, Billi AC, et al. (2012) MORC family ATPases required for heterochromatin condensation and gene silencing. Science 336: 1448-1451.

[2] Grob S, Schmid MW, Grossniklaus U (2014) Hi-C analysis in *Arabidopsis* identifies the *KNOT*, a structure with similarities to the *flamenco* locus of *Drosophila*. Mol Cell 55: 678-693.

[3] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

[4] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.

[5] Liao Y, Smyth GK, Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 41: e108.

[6] Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, et al. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods 9: 999-1003.

[7] Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, et al. (2007) Genome-wide profiling and analysis of *Arabidopsis* siRNAs. PLoS Biol 5: e57.

[8] Gregory BD, O'Malley RC, Lister R, Urich MA, Tonti-Filippini J, et al. (2008) A link between RNA metabolism and silencing affecting *Arabidopsis* development. Dev Cell 14: 854-866.

[9] Lister R, O'Malley RC, Tonti-Filippini J, Gregory BG, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133: 523-536.

[10] Jacob Y, Stroud H, LeBlanc C, Feng S, L Z, et al. (2010) Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. Nature 466: 987-991.

[11] Luo C, Sidote DJ, Zhang Y, Kerstetter RA, Michael TP, et al. (2013) Integrative analysis of chromatin states in *Arabidopsis* identified potential regulatory mechanisms for natural antisense transcript production. Plant J 73: 77-90.

[12] Stroud H, Greenberg MVC, Feng S, Bernatavichute YV, Jacobsen SE (2013) Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. Cell 152: 352-364.

7 Epilogue

In chapter 3 ("Polarized distribution of mRNA in the syncytial female gametophyte of *Arabidopsis thaliana* precedes cellularization and cell specification") we concluded that polarized localization of mRNA during syncytial development of the female gametophyte (FG) of *Arabidopsis thaliana* precedes cell fate decisions upon cellularization. We also noted that it remains to be clarified, whether these mRNAs directly act as cell-fate determinants, allow a rapid translational burst after cellularization, or if subcellular localization serves as a mechanism to simply control protein localization (e.g., to control the deposition of membrane proteins). However, we already pointed out that clarifying this question will require a series of challenging experiments, for some of which the technology may not yet be suitable (chapter 2). Once established and optimized, fluorescent *in situ* RNA sequencing (FISSEQ[1]) and MALDI-imaging mass spectrometry (MSI[2]) could allow for accurate subcellular localization of all transcripts and their proteins (see chapter 2 for a discussion of these methods). Further important questions regarding the polarized localization of mRNA in the syncytial FG may be adressed with a system facilitating *in vivo* monitoring of the subcellular localization of RNA in plants[3]. The system relies on fluorescent proteins fused to an RNA-binding protein specifically recognizing RNA stem-loops, which are fused to the mRNA under investigation. Tracking specific mRNAs *in vivo* would for example clarify whether the transcript-gradients are actively established/maintained, and if yes, the specific subcellular localization may guide future experiments to find the underlying machinery. I initiated these experiments during my thesis and cloned several candidate genes into a vector facilitating such an analysis. However, further experimental work is necessary to generate and supertransform the transgenic plant lines containing the λN_{22}-fluorescent protein fusions[3].

In summary, I believe that my work on the female gametophyte of *A. thaliana* provides a useful basis to identify novel genes involved in female gametogenesis, and may also aid to study the processes involved in intracellular mRNA transport in plants. Furthermore, the research carried out during my PhD thesis for my own and the collaborative projects revealed current challenges and potential future bottlenecks associated with modern quantitative biology. As novel high-throughput and large-scale methods are rapidly developed, the requirements for developing data analysis tools and approaches are currently not saturated and will further increase. Especially the use of cutting-edge technologies requires not only solid biological knowledge, but a substantial amount of computational and analytical skills. Therefore, analytical, statistical, and computational skills are and will be instrumental for molecular biology research, as exemplified in this thesis.

[1]Lee, J, *et al.* (2014) Science 343: 1360–1363.
[2]Schober, Y, *et al.* (2012) Analytical Chemistry 84: 6293–6297.
[3]Schönberger, J, *et al.* (2012) The Plant Journal 71: 171–181.

8 Appendix: Further contributions

8.1 Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture

The following manuscript is published in "Genome Biology" (open access)[1]. I designed the data handling concepts and implemented the raw data preprocessing. Stefan Grob and I designed the data analysis concepts together, and I implemented a part of it. I further contributed to data analysis and interpretation, wrote a part of the methods section, and helped to improve the manuscript.

[1] Grob, S, Schmid, MW, and Grossniklaus, U (2013) Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture. Genome Biology 14: R129.

Grob *et al. Genome Biology* 2013, **14**:R129
http://genomebiology.com/2013/14/11/R129

Genome Biology

RESEARCH Open Access

Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture

Stefan Grob[1], Marc W Schmid[1], Nathan W Luedtke[2], Thomas Wicker[1] and Ueli Grossniklaus[1*]

Abstract

Background: The packaging of long chromatin fibers in the nucleus poses a major challenge, as it must fulfill both physical and functional requirements. Until recently, insights into the chromosomal architecture of plants were mainly provided by cytogenetic studies. Complementary to these analyses, chromosome conformation capture technologies promise to refine and improve our view on chromosomal architecture and to provide a more generalized description of nuclear organization.

Results: Employing circular chromosome conformation capture, this study describes chromosomal architecture in *Arabidopsis* nuclei from a genome-wide perspective. Surprisingly, the linear organization of chromosomes is reflected in the genome-wide interactome. In addition, we study the interplay of the interactome and epigenetic marks and report that the heterochromatic knob on the short arm of chromosome 4 maintains a pericentromere-like interaction profile and interactome despite its euchromatic surrounding.

Conclusion: Despite the extreme condensation that is necessary to pack the chromosomes into the nucleus, the *Arabidopsis* genome appears to be packed in a predictive manner, according to the following criteria: heterochromatin and euchromatin represent two distinct interactomes; interactions between chromosomes correlate with the linear position on the chromosome arm; and distal chromosome regions have a higher potential to interact with other chromosomes.

Background

In eukaryotic nuclei, chromosomes of considerable length are densely packed into a very small volume. In *Arabidopsis*, chromatin with a total length of about 8 cm has to be packaged into a nucleus of about 70 μm^3 volume and 5 μm diameter [1,2]. Nonetheless, the extremely dense packaging of chromatin does not lead to a chaotic entanglement of chromatin fibers. Eukaryotes have evolved mechanisms to untangle chromatin and to organize the nucleus into structural domains, facilitating chromosome packaging and, hence, the accessibility of the information stored within chromosomes. Therefore, chromosomal architecture is likely to influence the transcriptional state of a given cell, and might be a major player in the epigenetic regulation of cell fate.

Over the past 15 years, the field of epigenetics has grown rapidly, addressing basic questions about the long-term regulation of genes, and how diverse cell types reach their differentiated states. These studies have provided insights into the mechanisms that enable cells to differentiate into diverse cell types with distinct phenotypes, despite sharing exactly the same genotype.

To date, most of the commonly studied epigenetic processes have been shown to involve covalent modifications of DNA, such as cytosine methylation, modifications of the core histone proteins H3 and H4, and histone variants. Thereby, chromatin can be grouped into activating and repressive chromatin states, defined by their epigenetic landscape. Among the main players are trimethylation of lysine 36 of H3 (H3K36me3) and dimethylation of lysine 4 of H3 (H3K4me2), which act as activating marks, and monomethylation of lysine 27 of H3 (H3K27me1) and dimethylation of lysine 9 of H3 (H3K9me2), which are associated with the repressive state [3-5].

* Correspondence: grossnik@botinst.uzh.ch
[1]Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, CH-8008 Zürich, Switzerland
Full list of author information is available at the end of the article

Although studied for over 100 years [6] (for example, with respect to cell division), chromosomal architecture, and thus higher-order chromatin organization, has not been a major focus of epigenetic research. Until recently, the lack of high-resolution techniques made structural studies of the nucleus extremely difficult. Nevertheless, chromatin condensation as seen in heterochromatin, reflecting, chromosomal architecture, could be viewed as the first described epigenetic mark [7,8]. Recently, it became possible to study chromosomal architecture in more detail, on both a global and a local scale, for instance with respect to physical interactions between enhancers and promoters [9,10].

In plants, chromosomal architecture has been studied for many years using cytogenetic techniques and microscopic observations. Early studies allowed the discovery of the basic chromosome conformations, heterochromatin and euchromatin, which were first described in mosses by Emil Heitz as early as 1929 [7]. Most condensed chromatin, or heterochromatin, is associated with centromeric regions. However, large heterochromatic regions outside the pericentromeres were also detected and, because of their microscopic appearance, were termed 'knobs'. Although first observed and best described in maize [11], knobs were also shown to exist in the model plant *Arabidopsis*, on chromosomes 4 and 5 [12-14]. The heterochromatic knob on the short arm of chromosome 4 (*hk4s*) is derived from an inversion event, which caused a pericentromeric region to lie in a more centrally located region of the chromosome arm. Owing to its length of 750 kb, *hk4s* is easily detectable, and is therefore the best studied knob in *Arabidopsis*. By contrast, the merely 60 kb long knob on chromosome 5 is only poorly described. Despite its central, and therefore euchromatic, position on the chromosome arm, *hk4s* has kept the heterochromatic features of its pericentromeric origin. The knob *h4ks* is characterized by low gene density and an abundance of highly repetitive sequences, such as transposable elements.

To date, two methods have been frequently used to study chromosomal architecture. For microscopic observations, fluorescence *in situ* hybridization (FISH) visualizes chromosomal architecture by detecting specific sections of chromosomes through hybridization with fluorescently labeled probes. Over the past decade, a completely different set of methods has been developed, which are summarized as chromosome conformation capture (abbreviated to 3C) technologies [15,16]. 3C uses formaldehyde cross-linked chromatin that is subsequently digested and religated. This produces circular DNA, comprised of two restriction fragments that were initially in close spatial proximity within the nucleus. The abundance of these circular 3C templates can then be used to calculate interaction frequencies between two given fragments in the genome. In both animal model systems and yeast, various studies have successfully used 3C technologies since the first publication in 2002 [15]. Whereas 3C is used to analyze pair-wise interactions (one specific fragment interacting with another specific fragment; that is, one to one), circular chromosome conformation capture (4C) identifies interactions genome-wide to a viewpoint of interest [17] (that is, one to all). HiC, the most recent 3C technology, facilitates the analysis of genome-wide interactions from all restriction fragments of a genome (that is, all to all) [18].

In the plant field, however, the adoption of these technical advances has been slower, and only a few studies have been performed using 3C technology. A 3C study in maize revealed chromatin looping at the paramutagenic *b1* locus [19], and another recent study showed the importance of local DNA looping for the correct expression of the flowering time regulator locus FLC [20]. Moissiard and colleagues compared global changes in the interactome between mutant *atmorc6* and wild-type plants [21]. However, that study did not focus on a detailed description of the chromosomal architecture of *Arabidopsis* nuclei.

Here, we provide insights into the general architecture of the *Arabidopsis* nucleus, using 4C applied to several viewpoints followed by Illumina sequencing. Our study aimed at characterizing global principles of chromosomal interactions and their correlations with epigenetic marks. Additionally, we found that the heterochromatic knob *hk4s* is characterized by a distinct interactome, which strongly resembles its pericentromeric origin.

Results

The current knowledge on chromosomal architecture in *Arabidopsis* is largely based on microscopic observations. Therefore, we aimed to gain insights into higher-order chromatin organization based on 4C technology, which promises to complement previously published FISH experiments, and to reveal novel mechanisms governing chromosomal architecture.

We performed 4C experiments on aerial tissue of 2-week-old *Arabidopsis* seedlings using thirteen specific restriction fragments (viewpoints) distributed across all five chromosomes (Figure 1A). Employing high-throughput sequencing, 4C technology identifies sequences that physically interact with a given viewpoint. Therefore, the position and number of mapped 4C sequencing reads define the interactome of the given restriction fragment (that is, the viewpoint) in space (position) and in frequency or specificity (number of reads).

To cover a wide distribution of chromosomal interactions, we chose viewpoints that reside in various locations: from pericentromeric, to mid-chromosome arm, to distal positions (Figure 1A).

Data evaluation reveals robustness of 4C experiments

To obtain the interactome of a given viewpoint, short sequence reads were mapped to restriction fragments,

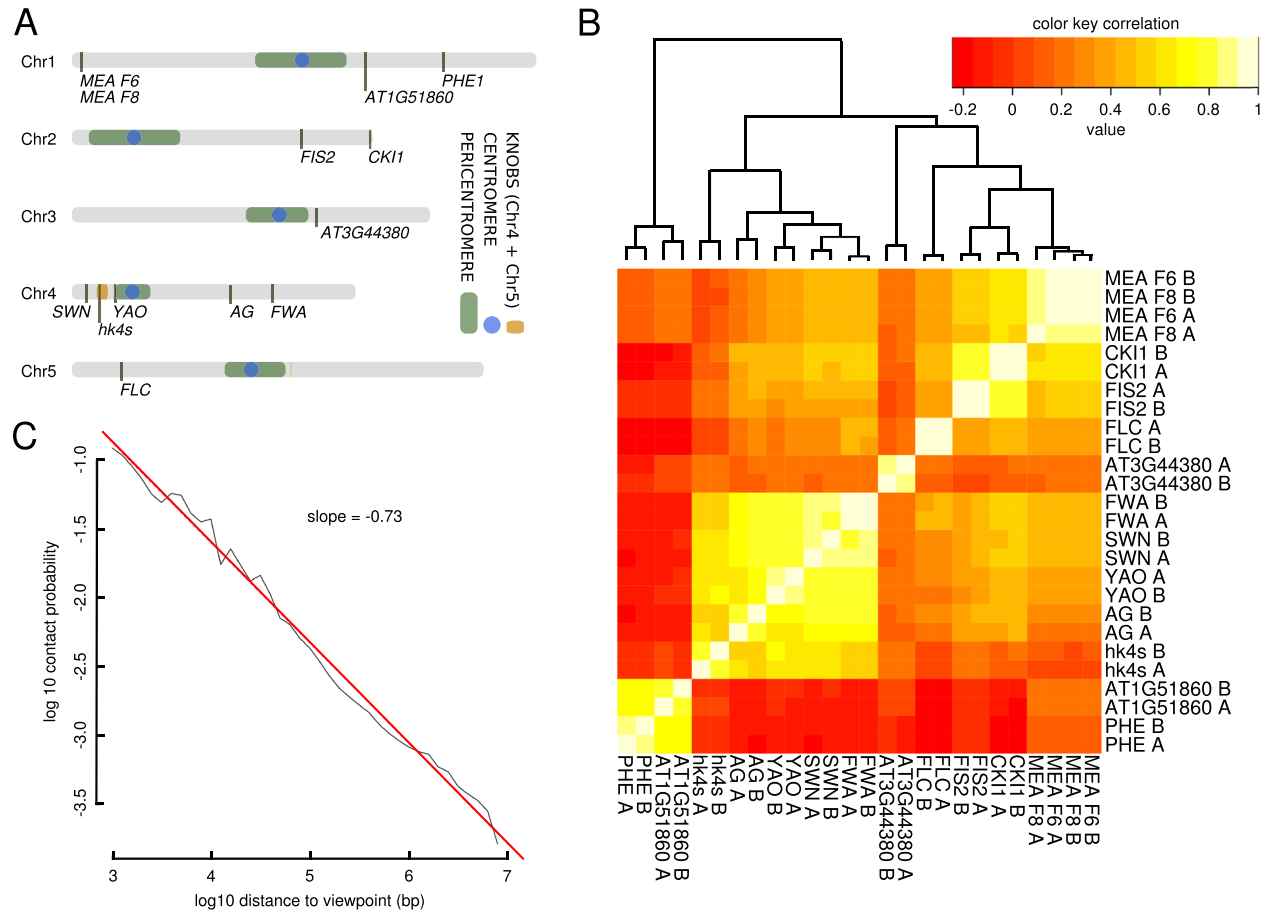


Figure 1 Primary circular chromosome conformation capture (4C) data analysis. (A) Schematic representation of the viewpoints chosen for this study. Viewpoints were named according to nearby genes or according to a region of special interest (*hk4s*). **(B)** Cluster analysis representing the reproducibility of biological duplicates. The letters 'A' and 'B 'at the end of the names indicate biological replicates. **(C)** Power law scaling, indicative of the interaction decay for all viewpoints, across a distance to the viewpoint from 1 kb to 10 Mb.

and subsequently merged into sliding windows consisting of 100 *Hind*III restriction fragments. We then assigned *P*-values to each window describing the specificity of the interaction to a given viewpoint. To obtain these *P*-values, read counts of 4C windows were compared with the probabilities of a normal distribution. The parameters of this distribution were calculated using 1,000 sets of windows, each generated by random shuffling of 4C fragments. As chromosome arms differ considerably in their length and, therefore, their DNA amount, we calculated *P*-values individually for each chromosome arm. Windows with $P \leq 0.01$ where defined as specifically interacting with their corresponding viewpoint and are, hereafter, referred to as 'preys'.

The mappability of sequencing reads poses a major concern for any genomic study. Owing to the incomplete assembly of centromeric repeats in the *Arabidopsis* reference genome, we excluded regions within 100 kb distance of the centromere. Visual inspection of genomic Illumina sequencing data revealed an even distribution of mapped reads along the remaining chromosome sequence and, therefore, no other major mappability biases were identified.

To assure the reproducibility of this study, 4C experiments were performed in duplicate. Correlations between duplicates and different viewpoints were calculated using the sum of reads per window. Spearman correlation coefficients were high for duplicates (mean ± SD 0.88 ± 0.07), and relatively low for different viewpoints (0.26 ± 0.31). However, interacting viewpoints and viewpoints located in close proximity (see Figure 1A), such as the two viewpoints at the *MEDEA* (*MEA*) locus, had correlation coefficients close to those of replicates of the same viewpoint. Cluster analysis supported these findings (Figure 1B), further demonstrating that viewpoints on the same chromosome arm also show higher correlations with each other than with viewpoints located on other chromosomes arms. Taken together, these analyses reveal the robustness of our data.

To differentiate between random interactions, which are mainly dependent on chromosomal proximity to

the viewpoint, and specific interactions, we estimated the genomic distance-dependent decay of the interaction probability on a distance of 1 kb to 10 Mb from the viewpoint. For this, we pooled 4C reads of all viewpoints within the given distance to their viewpoints. Performing linear regression on logarithmized distance and contact probabilities, we calculated a slope of −0.73, that is, the contact probability decays with a power law function of distance$^{-0.73}$ (Figure 1C). This result resembles similar analyses of the *Drosophila* (−0.85) [22] and human (−1.08) [18] genomes.

Cis interactions are enriched within chromosome arms

Because the replicate correlation was high, we pooled replicates for a common representation of the 4C interactome (Figure 2A,B) using the software Circos [23]. Figure 2C illustrates an example of a more detailed representation of 4C interactomes for the *FIS2* viewpoint. All other representations of individual viewpoints are shown in the additional files (see Additional file 1: Figure S1; Additional file 2: Figure S2; Additional file 3: Figure S3; Additional file 4: Figure S4; Additional file 5: Figure S5; Additional file 6: Figure S6; Additional file 7: Figure S7; Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10; Additional file 11: Figure S11; Additional file 12: Figure S12; Additional file 13: Figure S13). At first sight, we observed an apparent enrichment in inter-chromosomal interactions of distal regions of chromosomes (Figure 2A). Additionally, intra-chromosomal interactions appeared to be occurring mostly locally around the viewpoint and between the distal regions of the two chromosome arms (Figure 2B and Figure 2C).

Interactions can be categorized into *cis* and *trans* interactions, which require different analysis techniques [24]. *Cis* interactions (Figure 2B) refer to intra-chromosome interactions, whereas *trans* interactions (Figure 2A) are defined as inter-chromosome interactions.

By visual inspection of the interaction frequencies, we observed that local interactions rarely spread across the centromeres, (Figure 2B, Figure 2C; see Additional file 1: Figure S1; Additional file 2: Figure S2; Additional file 3: Figure S3; Additional file 4: Figure S4; Additional file 5: Figure S5; Additional file 6: Figure S6; Additional file 7: Figure S7; Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10; Additional file 11: Figure S11; Additional file 12: Figure S12; Additional file 13: Figure S13), indicating that interactions between the two arms of the same chromosome (that is, the inter-arm interactions) are distinct from the intra-arm interactions, thus splitting the *cis* interactions into two groups.

Therefore, we investigated whether chromosomes, or rather chromosome arms, are the basic unit of nuclear architecture. To answer this question, we calculated the average number of reads per million (RPM) for each chromosome arm, and defined three chromosome arm types: The chromosome arm hosting the viewpoint (viewpoint arm), the other arm on the same chromosome as the viewpoint (*cis* arm), and arms of all other chromosomes (*trans* arms). We observed the highest interaction frequencies and, therefore, the highest mean RPM values within the viewpoint arm (Figure 3A), showing that a high proportion of chromosomal interactions occur within the same arm.

Interactions with *cis* arms were significantly more frequent than those with *trans* arms (Student's *t*-test, $P = 0.0135$ for replicate A and $P = 0.0129$ for replicate B). However, the differences were small compared with the RPM values for the viewpoint arm and the *cis* arm (Student's *t*-test, $P = 1.4 \times 10^{-13}$ for replicate A and $P = 1.7 \times 10^{-13}$ for replicate B) (Figure 3A). A large proportion of interactions within the viewpoint arm occurred within the close vicinity of the viewpoint itself. To investigate whether long-range interactions also preferentially occur within the viewpoint arm, we excluded regions surrounding the viewpoints by 2 Mb on each side of the viewpoint (Figure 2A). Devoid of the viewpoint region, the RPM values were strongly reduced; however, they were still significantly higher than those of the *cis* arms (Student's *t*-test, $P = 0.012$ for replicate A and $P = 0.010$ for replicate B).

The difference between the *trans* and *cis* arms appears to be dependent on the distance of the viewpoint from the centromere. Distal viewpoints (for example, *MEA* and *CYTOKININ-INDEPENDENT1* (*CKI1*), see Additional file 1: Figure S1; Additional file 2: Figure S2; Additional file 6: Figure S6) did not appear to interact preferentially with their respective *cis* arm compared with the *trans* arm. This could been observed by comparing the overall interaction values of the viewpoint's respective *cis* arm compared with the overall interaction values of the *trans* arms. By contrast, viewpoints residing in the vicinity of the centromeres (for example, *YAOZHE* (*YAO*) and *AT3G44380*; see Additional file 7: Figure S7; Additional file 10: Figure S10) exhibited increased *cis* arm interactions compared with *trans* arm interactions and, thus, limited spreading of local interactions across the centromere.

In summary, intra-arm interactions were about ten-fold more frequent than inter-arm interactions, whereas inter-arm and inter-chromosomal interactions differed by about two-fold on average. Therefore, our results show that chromosome arms are the main interaction unit, and that interaction frequencies decrease sharply close to the centromeres.

Linear position along the chromosome influences the interaction potential of the viewpoint

We found that *trans* interactions could make up to 50% of the total interactome of a given viewpoint. Therefore, we were interested in understanding the mechanisms

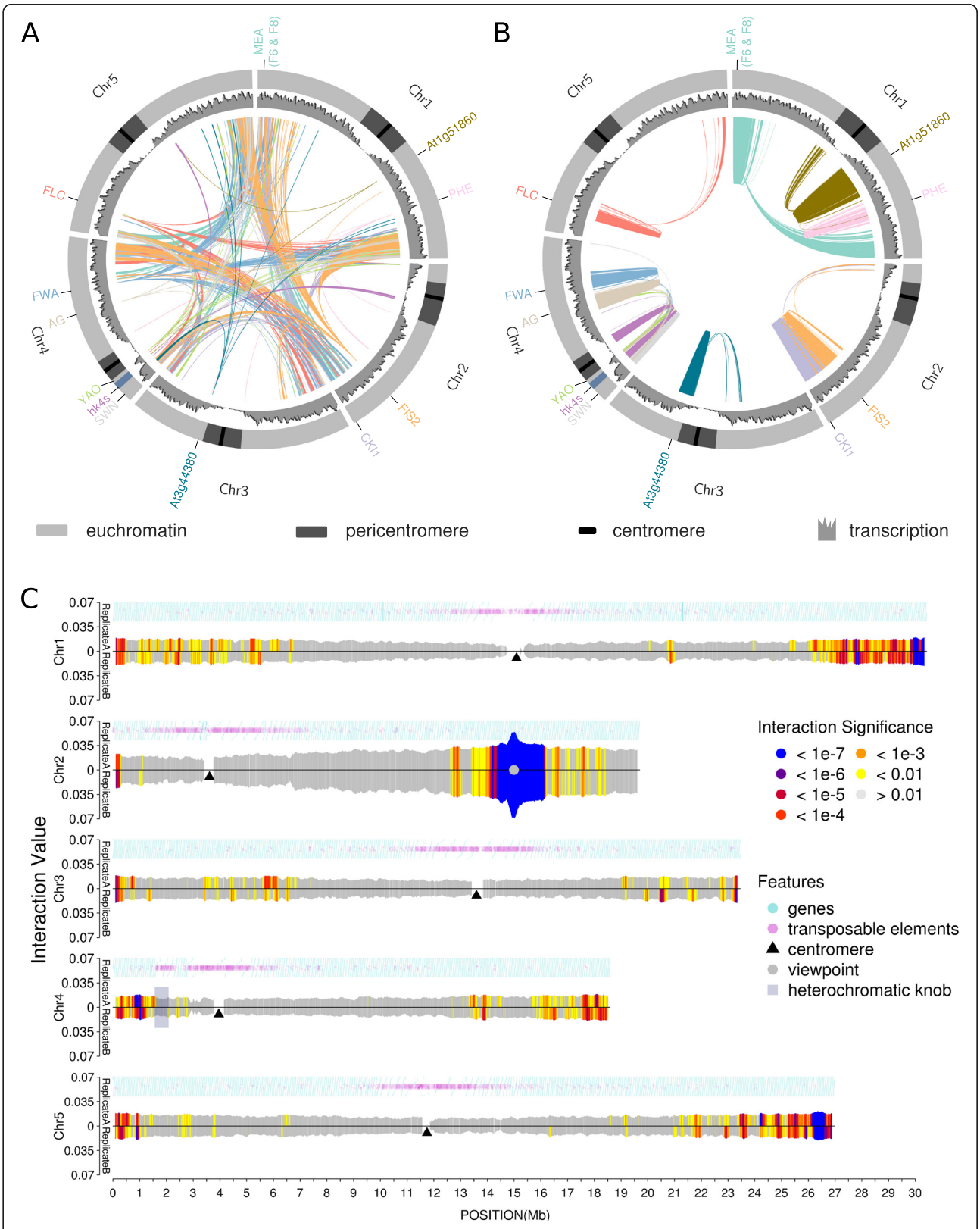


Figure 2 (See legend on next page.)

(See figure on previous page.)
Figure 2 Summary of circular chromosome conformation capture (4C) interactomes. Circos plots illustrate the 4C interactome, transcription rate, and chromosomes with euchromatic and centromeric regions. Line color refers to the color of the viewpoint names at the periphery of the Circos plots. Only interactions with a $P < 10^{-3}$ are plotted. **(A)** *Trans*- interactions; **(B)** *cis* interactions; **(C)** 4C interactome of viewpoint *FIS2*. Color code refers to significance levels. Gene density (blue circles) and transposable element density (purple circles) are indicated to illustrate the occurrence of heterochromatin and euchromatin. The region covered by the knob *hk4s* is highlighted with a transparent rectangle on the short arm of chromosome 4. Interaction values equal to $\sum_i(\log_2(\text{number of reads in fragment}_i))$, where i stands for a fragment within a given window, are scaled to the viewpoint's total library size.

governing *trans* interactions. Visual inspection of 4C data (Figure 2A, Figure 2C; see Additional file 1: Figure S1; Additional file 2: Figure S2; Additional file 3: Figure S3; Additional file 4: Figure S4; Additional file 5: Figure S5; Additional file 6: Figure S6; Additional file 7: Figure S7; Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10; Additional file 11: Figure S11; Additional file 12: Figure S12; Additional file 13: Figure S13) suggested an effect of the viewpoint positions along the chromosome arms on the *trans* interaction frequencies. We hypothesized that chromosomal interactions do not solely reflect specific functions of a given region, but are rather a consequence of physical constraints. To investigate whether the positioning of the viewpoints along the chromosome arm is a major constraint for *trans* interactions, we tested whether regions with similar distance to the centromeres are more likely to interact.

We calculated the relative distance to the centromeres, where 50% ($dist_{0.5}$) of all 4C reads could be found. As a considerable proportion of all interactions could be found surrounding the viewpoint and would therefore distort the analysis, we excluded the viewpoint arm. A significant correlation between $dist_{0.5}$ and the relative distance of the viewpoint to the centromere could be observed (Spearman correlation coefficient = 0.722; linear model $P = 3.4 \times 10^{-28}$) (Figure 3B). This suggests that regions with a similar relative distance to their corresponding centromeres are likely to co-localize with each other in the three-dimensional space of the nucleus. This observation was most pronounced in distal regions; however, it was also observable in regions in proximity to the pericentromeres.

Distal chromosomal regions show an increased *trans* interaction potential

We hypothesized that the flexibility of a chromosome arm is a major physical constraint influencing the interaction potential of a viewpoint. Assuming that centromeres act as chromosomal anchors, distal regions of chromosome arms should exhibit a higher flexibility than regions close to the centromere [25-28]. Hence, we predicted that distal viewpoints should exhibit an increased *trans* interaction potential.

Therefore, we tested the correlation between the absolute distance of the viewpoint to the centromere and the reads per kilobase per million (RPKM) of 4C reads found in *trans* (including the *cis* arm) (Figure 3C). Distal viewpoints were shown to interact more frequently with regions in *trans* than did viewpoints residing closer to the centromere (Spearman correlation coefficient = 0.774, linear model $P = 10^{-5}$) (Figure 3C).

These results indicate that the localization of a viewpoint along the chromosome arm significantly influences its interaction pattern.

Principal component analysis showed a correlation between the epigenetic landscape and the interactome

The interplay of epigenetic marks, such as histone modifications, and physical interactions of two sequences were previously shown to be important for stringent gene regulation [20,22,29,30]. Therefore, we investigated whether specific epigenetic marks can be correlated with long-range interactions.

We obtained previously published histone modification data [31], specifically H3K4me2, H3K4me3, H3K9me2, H3K27me1, H3K27me3, H3K36me2, H3K36me3, H3K9ac, and H3K18ac. From the same dataset, we included transcriptome, histone H3 occupancy, and genomic DNA control data. Additionally, we obtained publically available CG, CHH, and CHG DNA methylation data [32]. Because data obtained from chromatin immunoprecipitation (ChIP) for histone modifications cannot be directly compared with 4C data due to the different scaling of the two datasets [24], we calculated density values of each epigenetic feature within 4C windows. We analyzed the epigenetic modification densities (EMDs) as the sum of nucleotides covered by at least one uniquely alignable short sequence, divided by the total number of nucleotides for each individual 4C restriction fragment (that is, the length of the restriction fragment). Subsequently, the mean for each window was calculated. To adjust the scale of the 4C data to the EMDs, we chose a window size of 25 fragments, which still conferred satisfactory reproducibility between replicates. 4C windows were categorized into prey regions (windows that show an interaction probability of ≤0.01) and randomly chosen control regions.

If specific histone modifications or sets of histone modifications are associated with an interaction pair, it could be assumed that prey regions of a given viewpoint would share a common epigenetic environment, reflected by a particular composition of the EMDs. To elucidate how

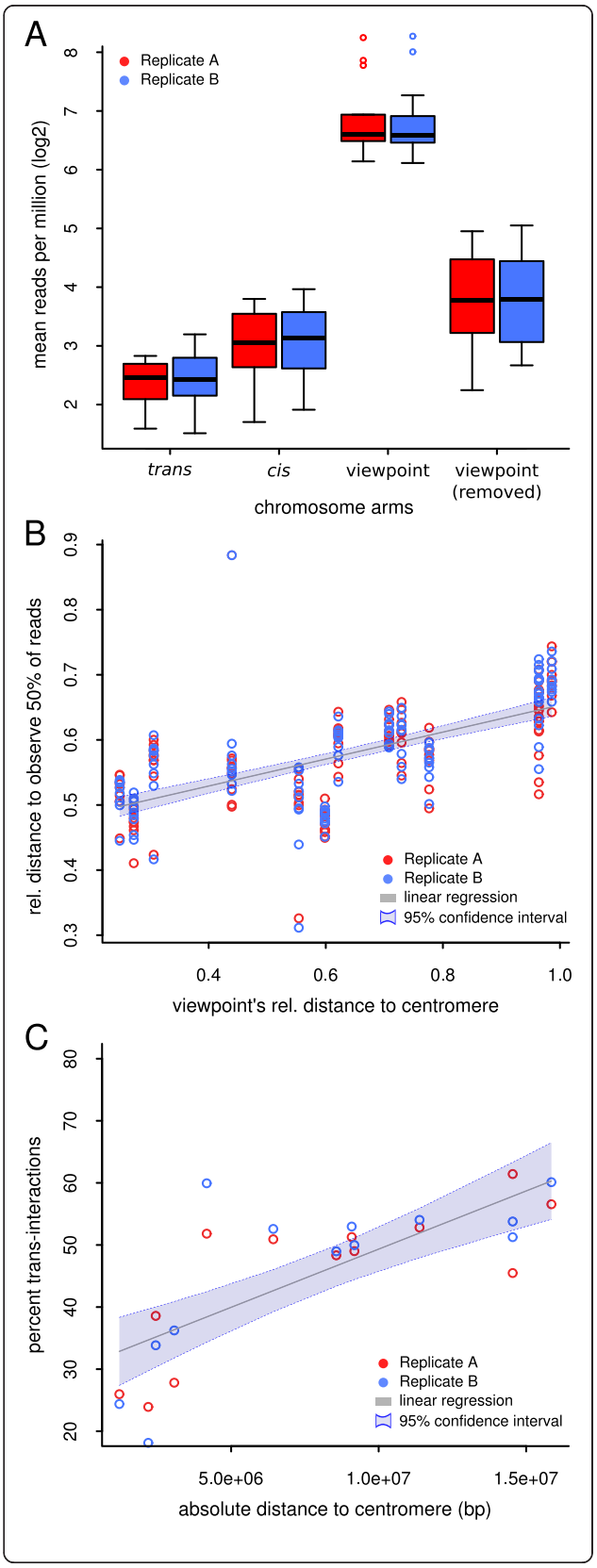


Figure 3 Physical constraints of chromosomal architecture. (A) Number of reads per million for four distinct classes of interactomes. Viewpoint: circular chromosome conformation capture (4C) reads that map on the same chromosome arm as the viewpoint. Viewpoint (removed): interactions mapping the viewpoint's arm, excluding interactions that map within 2 Mb distance on either side of the viewpoint. *Cis*: 4C reads that map to the other arm of the chromosome harboring the viewpoint. *Trans*: 4C reads that map to all other chromosome arms. **(B)** The relative distance to the centromere (0 at the centromere, 1 at the telomere) in which 50% of the 4C reads can be found depends on the relative distance of the viewpoint to the centromere. **(C)** The percentage of 4C reads that can be mapped to *trans* arms was positively correlated with the viewpoint's absolute distance to the centromere in base pairs (bp). In all parts, red circles represents replicate A, blue represents replicate B.

histone modifications are related to the interactome, we performed principal component analysis (PCA) (Figure 4A). For each viewpoint, the mean EMDs (selecting only histone modification data) of prey and control regions were calculated and included in the PCA. As the first principal component was found to explain 97% of the total variation, it was the only component used for further analyses.

Two opposing groups of EMDs, H3K36me3/H3K4me2 and H3K27me1/H3K9me2, were found to be the major contributors to the first principal component of the PCA (Figure 4A, arrows). Closer observation of three viewpoint/prey pairs revealed how EMDs and interaction frequencies are coupled (Figure 4C). Euchromatic viewpoints, such as *FLOWERING WAGENINGEN* (*FWA*) (Figure 4C, top row), which are characterized by low levels of H3K27me1 and enrichment of H3K36me3, preferentially interacted with regions of a similar EMD pattern. This is evident from the increased H3K36me3 levels surrounding the region of high interaction frequencies and local peaks of H3K27me1 enrichment, coinciding with a significant drop in interaction frequencies (Figure 4C, top row, right panel). By contrast, heterochromatic viewpoints (Figure 4C, middle and bottom rows), which are characterized by the inverse EMD composition, preferentially interacted with regions exhibiting low H3K36me3 and high H3K27me1 levels. For example, local enrichment of H3K27me1 coincided with increased interaction frequencies to *PHE1* (Figure 4C, middle row, right panel). Moreover, the asymmetric local interactions surrounding *hk4s* appeared to be reflected by the asymmetric distribution of H3K27me1 (Figure 4C, bottom row, left panel).

Additionally, we performed PCA separately for individual viewpoints (see Additional file 14: Figure S15). Although the same EMDs could be identified as major factors for most viewpoints, the first component of the PCA was less dominant, indicating a more complex collaboration of factors separating control regions from prey regions. Furthermore, various viewpoints did not show a very clear separation of prey and control regions.

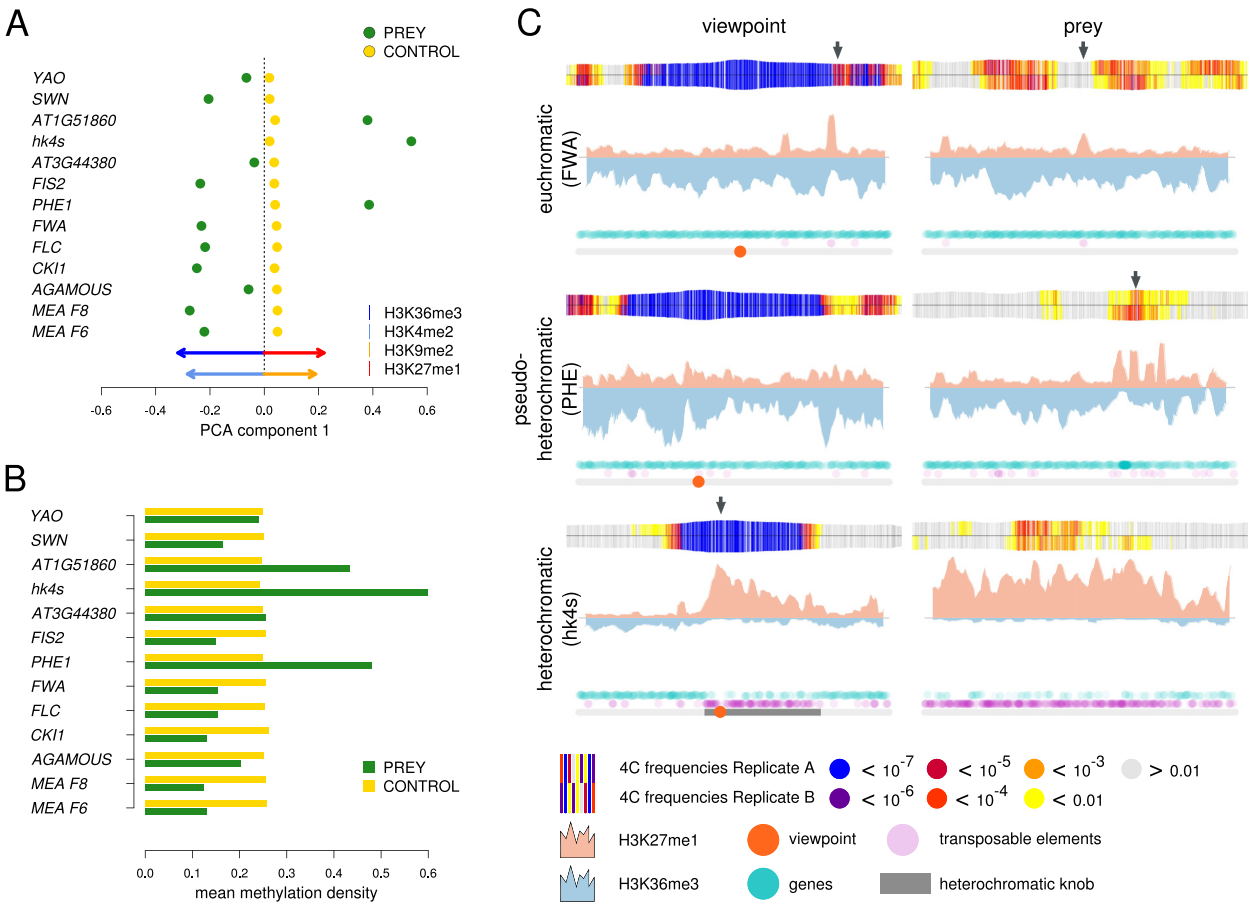


Figure 4 Crosstalk of epigenome and interactome. (A) Principal component analysis (PCA) using mean epigenetic modification densities (EMDs) of control and prey regions for each viewpoint. EMDs included in the PCA were: H3K4me2, H3K4me3, H3K9me2, H3K27me1, H3K27me3, H3K36me2, H3K36me3, H3K9ac, and H3K18ac. Colored arrows represent the two highest contributing EMDs to the variance of the first component in positive and negative direction, respectively. Note that the first principal component explains almost all the variance (97%), and therefore, this was the only component plotted. Prey regions are represented by green dots, control regions by yellow dots. **(B)** Mean CG methylation densities of prey and control regions for individual viewpoints. The mean was calculated across 1000 times randomly sampled 50 prey and 50 control regions, respectively. Green bars represent preys and yellow bars represent controls. **(C)** Examples of the interactome-epigenome interplay for three different viewpoints and one of their corresponding prey regions. Top track: log summed 4C reads per window (100 fragments, starting every fragment). 4C reads of replicate A are plotted in the positive intercept, and 4C reads of replicate B are plotted in the negative intercept. Middle Track: EMD of the highest contributing factors of the PCA in positive and negative direction, respectively. In order to achieve comparable representation of H3K36me3 and H3K27me1 densities, the density of every window (25 fragments, starting every 5 fragments) was divided by the mean density of each histone modification. Arrowheads point at regions where the 4C interactome and local EMD peaks appeared to correlate. *FWA*: viewpoint on chromosome 4, 12 to 14 Mb; prey on chromosome 5, 23 to 25 Mb. *PHE*: viewpoint on chromosome 1, 23.5 to 25.5 Mb; prey on chromosome 1, 20 to 22 Mb. *hk4s*: viewpoint on chromosome 4, 0.8 to 2.8 Mb; prey on chromosome 2, 4 to 6 Mb.

Interestingly, this was most evident for viewpoints whose preys are associated with heterochromatic marks (*PHERES1* (*PHE1*), *hk4s*, *AT1G51860*) (see Additional file 14: Figure S15).

To address the individual contribution of epigenetic marks to the interactome, we performed a test based on a modified Gene Set Enrichment Analysis (GSEA) [33]. In summary, we tested whether prey regions would show a non-random distribution in their EMD profiles (see Materials and Methods for a detailed description). The obtained empirical *P*-values are indicative of the likelihood of a random set of regions to show a similar distribution of EMD values as the tested prey regions (Table 1).

To independently investigate whether control and prey regions differ significantly for individual epigenetic features, we developed a permutation test. In the first step, we calculated for each viewpoint the mean density for each epigenetic feature (Figure 4B and Additional file 15: Figure S16). Epigenetic features that coincide with the occurrence of heterochromatin and euchromatin, such as DNA methylation, clearly split the viewpoints into two

Table 1 Analysis of the epigenetic landscape

Genomic feature	P-value[a]	
	Permutation test	GSEA-like test
H3	0.1013	0.0779
H3K18ac[b]	**0.0335**	**0.0178**
H3K27me1[b]	**0.0249**	**0.0084**
H3K27me3	0.3355	0.099
H3K36me2[b]	**0.0033**	**0.0051**
H3K36me3[b]	**0.0033**	**0.0054**
H3K4me2[b]	**0.0033**	**0.0051**
H3K4me3[b]	**0.0037**	**0.0051**
H3K9ac[b]	**0.0033**	**0.0051**
H3K9me2[b]	**0.0325**	**0.0057**
Transcription[b]	**0.0033**	**0.0054**
CG methylation replicate 1[b]	**0.0065**	**0.0054**
CHG methylation replicate 1[b]	**0.0083**	**0.0051**
CHH methylation replicate 1[b]	**0.0083**	**0.0051**
CG methylation replicate 2[b]	**0.0083**	**0.0054**
CHG methylation replicate 2[b]	**0.0087**	**0.0051**
CHH methylation replcate 2[b]	**0.0083**	**0.0051**
Genomic DNA	0.0871	0.056

[a]Table contains adjusted *P*-values (false discovery rate; FDR (Benjamini-Hochberg)) for genomic features tested with a permutation test or a Gene Set Enrichment Analysis (GSEA)-like algorithm.
[b]Genomic features differing significantly between prey and control regions ($\alpha = 0.05$).

groups. Whereas viewpoints such as *PHE1*, *AT1G51860*, and *hk4s* had high methylation levels in their prey regions and low methylation levels in control regions, viewpoints that occur in euchromatin showed an inverse pattern. Similar patterning was also detectable for other epigenetic modifications (Figure 4B; see Additional file 15: Figure S16).

The inverse patterning of the epigenetic landscape between different viewpoints made it difficult to perform statistical tests using EMD values directly. Therefore, we calculated the absolute difference in the density of the epigenetic features density between control and prey regions. In essence, we tested whether the absolute difference in EMD values between prey and control regions were significantly different from the absolute difference between two sets of randomly selected regions. As a test set, we shuffled the 50 prey and 50 control regions into two randomized groups. As for the prey and control regions, we then calculated means and subsequently absolute differences between the two randomized groups. By repeating the permutations 1,000 times, we obtained a distribution of absolute differences between the two randomized groups for each epigenetic feature. This allowed us to calculate empirical *P*-values, which describe the chance that two randomly selected regions would differ more in their EMD setup than would prey and control regions (Table 1).

In line with the previously performed PCA, both tests revealed that the densities of most epigenetic features differed significantly between control and prey regions (Table 1). Histone H3 occupancy, however, did not differ significantly between the two groups, indicating that histone density itself does not correlate with a viewpoint's interactome. Additionally, no significant difference in genomic control data could be observed, rendering possible sequencing and alignment biases of the analyzed EMD dataset unlikely.

In summary, we conclude that the epigenetic landscape coincides with the interactome. This is mainly reflected by distinct euchromatic and heterochromatic interactomes.

The heterochromatic knob evades its euchromatic environment

Analyzing the read numbers of a first set of 4C viewpoints, we consistently observed a drop in read numbers for a region situated in the center of the short arm of chromosome 4 (Figure 5B; see Additional file 1: Figure S1; Additional file 2: Figure S2; Additional file 3: Figure S3; Additional file 4: Figure S4; Additional file 5: Figure S5; Additional file 6: Figure S6; Additional file 7: Figure S7; Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10; Additional file 11: Figure S11; Additional file 12: Figure S12; Additional file 13: Figure S13). Unexpectedly, this drop in interaction frequency was observed irrespective of the location of the viewpoint. Additionally, we did not observe this drop with visual inspection of genomic sequencing data, implying no mappability bias. Therefore, we hypothesized that global constraints of chromosomal architecture govern genome-wide interactions with this region.

Exploring the region in more detail, we found that it corresponds to the heterochromatic knob (*hk4s*), which is cytogenetically detectable and has been described previously [12,34] (see Additional file 9: Figure S9).

To analyze the implications of *hk4s* on chromosomal architecture in more detail, we designed three additional 4C assays. We set a viewpoint within *hk4s* and two viewpoints flanking *hk4s* in a more distal region (*SWINGER* (*SWN*)) and a more proximal region (*YAO*) of the short arm of chromosome 4. As the flanking viewpoints were set relatively close to *hk4s*, we expected increased frequencies of interactions within the knob and the viewpoints, owing to the previously observed local enrichment of interactions surrounding the viewpoints. However, the local interaction frequency of both neighboring viewpoints dropped sharply on the borders of *hk4s* (Figure 5A, Figure 5B; see Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10). *YAO* (coordinate at 2.75 Mb) is situated adjacent to the border of the pericentromere (coordinates

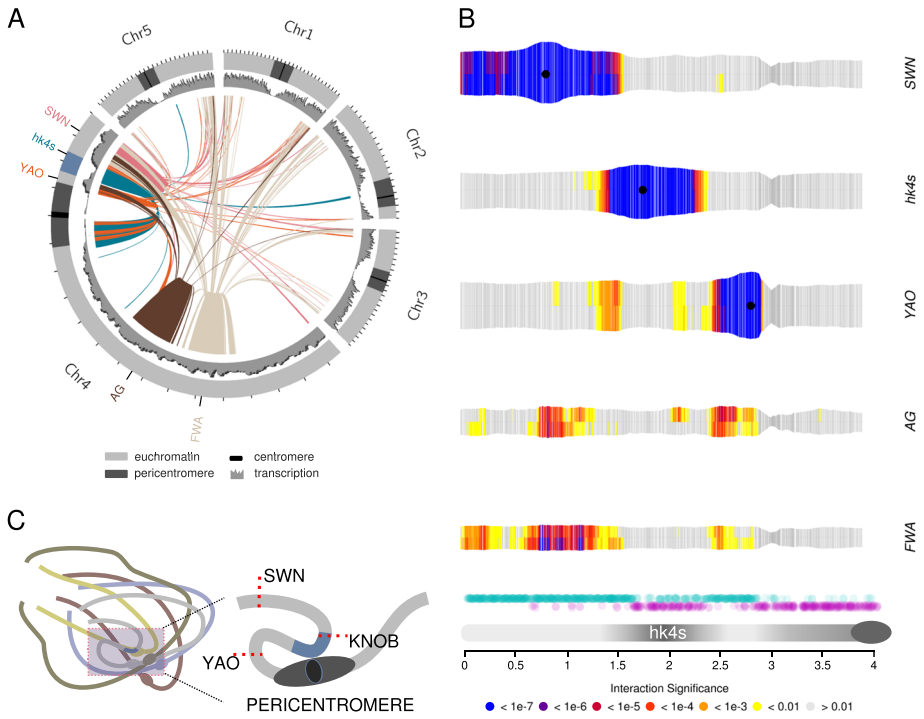


Figure 5 Interactome of the knob *hk4s*. (A) Circos plot illustrating all *cis* and *trans* interactions of viewpoints located on chromosome 4. Only interactions with $P \leq 10^{-4}$ were considered. Line color corresponds to the color of the viewpoints name indicated at the periphery of the plot. Chromosomes are not drawn to scale. **(B)** Representation of interaction frequencies for viewpoints situated on chromosome 4. Note that only the region up to 4 Mb is plotted, therefore, viewpoints *AG* and *FWA* cannot be seen. Black dots show positions of viewpoints; turquoise dots, genes; violet dots, transposable elements; light grey, euchromatic chromosomal segment; dark grey, heterochromatic chromosomal segments; dark grey ellipse, centromere. **(C)** Model of a potential mid-range chromosomal loop, connecting *hk4s* with the centromere of chromosome 4.

2.78 to 5.15 Mb) [3]. Interestingly, the local interaction pattern appears to be asymmetric. We observed a loss of specific interactions not only along the boundary to the knob but also along the much closer border of the pericentromeric region (Figure 5B; see Additional file 10: Figure S10). The defined sharp boundaries for local *YAO* interactions resembled the interaction pattern of *hk4s*. Whereas *YAO* resides in euchromatin surrounded by heterochromatin, *hk4s* can be viewed as its counterpart, residing in heterochromatin but surrounded by euchromatin (Figure 5B).

Regions situated on the long arm of chromosome 4 (*AGAMOUS* (*AG*) and *FWA*) interacted strongly with regions surrounding *hk4s,* including *YAO*, but not with *hk4s* itself (Figure 5B; see Additional files 11: Figure S11; Additional file 12: Figure S12), resembling the sharp drop in the interaction frequencies of *SWN* and *YAO* (Figure 5A, Figure 5B; see Additional file 8: Figure S8; Additional file 9: Figure S9; Additional file 10: Figure S10).

Consistent with observations for the two flanking viewpoints, the significant local interaction frequencies of the viewpoint set in the center of *hk4s* were limited by the borders of the knob. Additionally, we observed strong interactions of *hk4s* with the pericentromeric regions of chromosome 4 and with the pericentromeres of other chromosomes (Figure 5A). The apparent absence of specific interactions between *hk4s* and the pericentromere of the short arm of chromosome 4 is likely to be an artifact of the method used to assign *P*-values. Indeed, as *P*-values were calculated for individual chromosome arms, the high number of reads covering the viewpoint itself masks other regions on the same chromosome from being associated with low *P*-values.

Discussion

Replication and the choice of appropriate window size are key to ensuring robustness of 4C

Based on a correlation analysis of biological replicates, we show that 4C interaction profiles in *Arabidopsis* can be reproducibly obtained. However, reproducibility is dependent on the window size chosen. As chromosomal interactions are dynamic and partly stochastic, one single restriction fragment of two replicates can vary considerably in read number. Taking windows consisting of several fragments into account can balance this variation. As we were mainly interested in the global architecture of the *Arabidopsis* nucleus, we chose window sizes of up to 100 restriction fragments. However, the resolution for studying

short-range interactions is decreased by increasing the window size. Whereas 4C is well suited to study mid-range and long-range interactions in *Arabidopsis*, it is not necessarily the method of choice to study short-range interactions (for example, promoter/enhancer interactions). Regulatory sequences that are presumably involved in short-range interactions, such as chromatin loops, are often separated by less than a few kb. They are, therefore, difficult to analyze using 3C technologies, which rely on a sufficient number of restriction sites between the two regions of interest to confer satisfactory resolution.

Arabidopsis and *Drosophila* show comparable chromatin compaction and genome size

The interaction decay exponent describes the slope with which the interaction probability decays from the viewpoint. Therefore, it can provide an approximation of regional chromosomal compaction. Theoretically, a steeper slope indicates decreased flexibility of a given viewpoint, as distant regions are less likely to interact with it. Decreased flexibility can be interpreted as higher local chromatin compaction. *Drosophila* and *Arabidopsis* are similar with respect to chromosome number, genome size, total number of genes, and nuclear volume [1,35]. These characteristics could lead to similar constraints of chromosomal architecture. The interaction decay exponent determined in this study (–0.73) is close to that described earlier for *Drosophila* (–0.85) [22]. Interestingly, the interaction decay exponent in human nuclei is lower (–1.08), implying higher local compaction [18]. This observation is consistent with the physical characteristics of human nuclei compared with those in *Arabidopsis* and *Drosophila*. Although varying considerably, human nuclei show a lower volume/DNA ratio than the nuclei in *Drosophila* and *Arabidopsis*, indicating a higher global chromatin compaction [35]. It is important to mention, however, that interaction decay exponents cannot be compared very easily between different studies, as the calculated exponents of the power law scaling depend on the range of distances used for calculations. However, which scale best describes an overall distance-dependent interaction decay is a matter of debate. Additionally, the slope with which interactions decay was previously shown to vary between domains with different epigenetic landscapes [18,22]. We observed a variation in interaction decay exponents between the different viewpoints, from –0.56 to –0.96 (see Additional file 16: Figure S14). However, we could not explain these differences, either by the positional or by the epigenetic environment of a given viewpoint. Therefore, the global distance-dependent interaction decay does not necessarily add to the understanding of how interaction frequencies decrease with distance from an individual viewpoint.

How and whether global nuclear compaction and interaction probability decay really correlate is not entirely clear. An exploration of the *Arabidopsis linc1,linc2* double mutant could possibly answer this question, as these plants were reported to exhibit increased DNA density compared with wild-type plants [1].

4C results refine the view on general chromosomal architecture in *Arabidopsis*

The investigation of general features of chromosomal architecture in this study is consistent with previous findings studying *Arabidopsis* nuclei using cytogenetic methods [27,36]. However, 4C technology enables us to generate genome-wide interaction maps for various viewpoints and, hence, does not depend on a pair-wise analysis of two interacting sequences. This greatly adds to our understanding of general constraints on chromosomal architecture.

Basic interaction units appear to be defined as chromosome arms, with centromeres acting as a boundary. These findings are in agreement with an earlier study by Schubert and colleagues, reporting that chromosome arms are localized in distinct territories, as evidenced by FISH on *Arabidopsis* nuclei [36]. However, whether centromeres always act as strict boundaries cannot be conclusively answered, as the boundary effect of centromeres is likely to vary between the different chromosomes.

We observed a strong influence of the chromosomal location of a viewpoint on its interaction potential. Remarkably, the linear organization of chromosomes was reflected in the overall interaction potential of a given viewpoint, despite the dense packaging of the genome in the nucleus.

We propose that centromeres anchor the chromosomes in the nucleus, thereby allowing chromosome arms to protrude inside the nuclear volume [25-28]. The flexibility of chromosome arms thus increases with their length, allowing distant regions to interact more frequently in *trans* than more centrally located regions. Our hypothesis is supported by strong evidence for clustering of centromeres and their adherence to the nuclear matrix in different model organisms [37-39]. Taken together, these findings may explain why regions with a similar distance to the centromeres, which act as anchor points, preferentially interact with each other.

We also observed significant inter-telomeric interactions. A high interaction frequency of (sub-)telomeric regions in *Arabidopsis* was recently also shown by FISH [36]. In addition, previously published HiC data suggest increased interaction frequencies between telomeres [21,38]. By contrast, telomeres and centromeres do not interact, indicating a strict separation of these two key organizational elements of *Arabidopsis* chromosomes. These findings are in line with previous studies, and may be explained by the nucleolar localization of telomeres [27,40].

Remarkably, in *Drosophila*, long-range interactions seem to occur nearly exclusively within the viewpoint's chromosomal arm [30]; however, in the present study, up to 50% of all interactions were found to be outside this region. Whether this difference from *Drosophila* holds biological meaning is unclear. The presence of a higher number of individual cell types in the sample could theoretically increase the number of observable interactions, and result in a more complex interactome of a given viewpoint. Such increased complexity could thereby lead to an increased number of *trans* interactions. However, we do not estimate the number of cell types to be significantly different between the present study and the report by Tolhuis and colleagues, in which 4C was performed on *Drosophila* larval brain tissue [30], as the aerial seedling tissue used in our study is predominantly composed of mesophyll cells. The phase of the cell cycle might be a more important confounding factor. Over a cell cycle, chromosomal architecture changes dramatically. Cells of *Arabidopsis* seedlings divide at high frequency, leading to a rather short time period in which cells reside in interphase. Therefore, the proportion of cells in specific stages of the cell cycle could be a major factor influencing the (average) chromosomal conformation of a population of cells.

The interactome of a viewpoint is reflected in its epigenetic landscape

PCA revealed two distinct groups of prey regions, which could be discriminated mainly by the level of H3K36me3/H3K4me2 and H3K27me1/H3K9me2 densities. Interestingly, these histone modifications are commonly attributed to euchromatin or heterochromatin, respectively [31]. Furthermore, the heterochromatic pair H3K27me1/H3K9me2 is described to be the major component of 'chromatin state 3', which is mainly associated with transposable elements, as previously reported by Roudier and colleagues, whereas the pair H3K36me3/H3K4me2 primarily contributes to 'chromatin state 1', associated with active genes [3]. Filion and colleagues describe five distinct chromatin types in *Drosophila*, distinguished by the composition of proteins adhering to the DNA. H3K4me2 was shown to be most abundant in 'red chromatin', which represents one of two euchromatic chromatin states, whereas H3K9me2 is enriched in 'green chromatin', which can best be described as the classic heterochromatin of pericentromeric regions [4]. As anticipated by previous cytological studies of *Arabidopsis* nuclei, the interactome obtained by 3C technologies can be separated into two distinct domains, correlating with both the epigenetic and the cytogenetic definition of heterochromatin and euchromatin. Interestingly, this distinction is not only confined to *cis* interactions but can also be observed at the level of the whole genome. In addition, we suggest a further discrimination of heterochromatic interactions. The purely heterochromatic viewpoint *hk4s* predominantly interacts with visible heterochromatin such as the pericentromeric regions. *PHE1*, which shows moderate H3K27me1 enrichment surrounding the viewpoint, interacts predominantly with heterochromatic islands within otherwise euchromatic regions (Figure 2, Figure 4C; see Additional file 4: Figure S4).

Previous work in *Arabidopsis* has shown that homologous pairing is decreased in hypomethylation mutants [41], indicating a role for cytosine methylation in long-range interactions. We observed significant differences between control and prey regions with respect to their CG, CHH, and CHG methylation densities. Additionally, transcription rates exhibited significant differences between prey and control regions. Whether transcriptionally active genes interact with each other is not clear, as the genes residing in our viewpoints were not evenly balanced with regard to their transcriptional state (active versus silenced), rendering them inappropriate for statistical analysis.

Taking these results together, we conclude that interactomes share a common epigenetic landscape, leading to distinguishable heterochromatic and euchromatic interactomes. However, it is not clear to what extent individual epigenetic modifications influence the interactome, and to what extent the epigenetic landscape is the cause or consequence of a given interactome.

The knob *hk4s*: exception or rule?

Finally, the knob *hk4s* appears as an exceptional feature within the *Arabidopsis* nuclear landscape, as it interacts predominantly with pericentromeric regions. We think that *hk4s* represents the exception that proves the rule because its interactome reflects the pericentromeric origin of *hk4s*, which arose by an inversion that placed a pericentromeric region into the center of the chromosome arm. As discussed above, heterochromatic regions form a distinct interactome, in which heterochromatic islands that reside in an euchromatic environment are included. Figure 5C illustrates a model suggesting overall chromosomal architecture and chromosomal looping of *hk4s* to the clustered centromeres. Our results indicate that the knob *hk4s* acts as an interaction insulator for its neighboring regions, and conserves its pericentromeric origin with respect to its interaction frequencies.

To date, neither a functional role as a (neo)centromere nor an association with the nuclear matrix has been reported for *hk4s*. However, the specific interaction of *hk4s* with centromeres could raise speculation concerning the functional role of *hk4s* in the nucleus. The specificity of a given region to function as a centromere is surprisingly flexible. Previous reports show that in maize, centromere identity is not irreversibly defined. Wolfgruber and colleagues demonstrated that the centromere of maize chromosome 5 has moved to a new location, due to the invasion of non-centromeric retrotransposons, splitting the

centromere into two. Consequently, one of the two cleavage products lost its association with histone CenH3, which defines centromeres epigenetically by replacing the regular histone H3 protein [42]. In maize, centromere identity correlates with the abundance of centromeric retrotransposons [43], which specifically invade centromeric regions. Nevertheless, centromere identity appears to be mainly controlled epigenetically and not by DNA sequence [44,45]. However, previous reports show that that histone CenH3 accumulation defines the functional centromere in *Arabidopsis* and that CenH3 is predominantly associated with the 178 bp centromeric repeats [46,47]. As the knob *hk4s* lacks the centromeric 178 bp repeats and is thought to originate from a pericentromic region, which is not associated with CenH3, we conclude that *hk4s* is mainly involved in heterochromatin formation, and that *hk4s* is unlikely to play a role as a (neo)centromere.

Conclusions

Centromeres are key elements for chromosomal organization, as the position relative to the centromere strongly influences the interactome of a chromosomal region. We propose that the length of chromosome arms limits the mobility with which a region can traverse through the nuclear space and, therefore, influences the interaction potential in *trans*. Another hallmark of chromosomal architecture in *Arabidopsis* nuclei is the separation of two seemingly distinct interactomes, strongly correlating with visible heterochromatin and euchromatin. Interestingly, heterochromatic islands are partly able to evade their euchromatic context. The epigenetic landscapes of the heterochromatic and euchromatic interactome are clearly distinguishable. Therefore, histone modifications, which were previously described to be characteristic of chromatin states, may also be predictive for the interaction potential of a given chromosomal region.

Materials and methods

Nuclei extraction and 4C sample preparation

Seedlings of *Arabidopis thaliana* (L.) Heynh, accession Columbia (Col-0), were grown for 14 days on MS plates (4.3 g/l Murashige and Skoog salt (Carolina Biological Supply Company, Burlington, North Carolina, USA), 10 g/l sucrose (Applichem GmbH, Darmstadt, Germany), 7 g/l PHYTAGAR (Life Technologies Europe, Zug, Switzerland), pH5.6). Aerial tissue of seedlings was collected (approximately 10 g per sample), and distributed evenly between four conical 50 ml tubes. Under vacuum, the seedlings were incubated for 1 hour at room temperature in 15 ml freshly prepared nuclei isolation buffer (NIB: 20 mmol/l Hepes (pH8), 250 mmol/l sucrose, 1 mmol/l $MgCl_2$, 5 mmol/l KCl, 40% (v/v) glycerol, 0.25% (v/v) Triton X-100, 0.1 mmol/l phenylmethanesulfonylfluoride (PMSF), 0.1% (v/v) 2-mercaptoethanol) and 15 ml 4% formaldehyde solution, then 1.9 ml of 2 mol/l glycine was added to quench the formaldehyde, and the mixture was incubated for another 5 minutes under vacuum. The seedlings were snap-frozen in liquid nitrogen, and ground to a fine powder. The powder from two initial tubes was pooled and suspended in 10 ml NIB, with added protease inhibitor (Complete Protease Inhibitor Tablets; Roche, Basel, Switzerland; two tablets in 150 ml NIB). The suspension was filtered twice through Miracloth (Calbiochem/EMD Milipore, Darmstadt, Germany) adding an additional 10 ml NIB. The filtered nuclei suspension was spun for 15 minutes at 4°C and 3000×*g*. The supernatant was discarded, and the pellet was resuspended in 4 ml NIB and transferred to two 1.5 ml reaction tubes. After the tubes were spun for 5 minutes at 4°C and 1900×*g*, the supernatant was removed, and the pellet was resuspended in 1 ml NIB, followed by centrifugation under the above conditions. This step was repeated twice. Then, the nuclei were washed twice with 1.2 × NEB buffer 4 (New England Biolabs, Ipswich, MA, USA) (10 × NEB buffer 4: 50 mmol/l potassium acetate, 20 mmol/l Tris acetate, 10 mmol/l magnesium acetate, 1 mmol/l dithiothreitol (DTT)), using the centrifugation conditions described above. The nuclei were finally resuspended in 500 ml 1.2 × NEB buffer 4, with 5 μl of 20% SDS added. The samples were incubated for 40 minutes at 65°C, followed by 20 minutes at 37°C under constant shaking, then 50 μl of 20% Triton X-100 were added. The mixture was incubated for 1 hour at 37°C under constant shaking, then 60 μl of sample was removed as a pre-digestion control.

For digestion 15 μl 10 × NEB buffer 4 and 115 μl H_20 were added to the samples, and digestion was started using 100 U of *Hind*III restriction enzyme (New England Biolabs). After 3 hours of incubation at 37°C, 200 U of *Hind*III were added, followed by overnight incubation at 37°C. Next morning 100 U of *Hind*III were added, and samples were incubated for a final 2 hours. An aliquot (80 μl) of the sample was transferred to a fresh tube, and kept aside as a post-digestion control. To inactivate *Hind*III, 20 μl 20% SDS were added, and samples were incubated at 65°C for 25 minutes under constant shaking. Samples were transferred to 15 ml conical tubes, and 700 μl of 10× ligation buffer (0.5 mol/l Tris-Cl, 0.1 mol/l $MgCl_2$, 0.1 mol/l DTT, pH 7.5), 375 μl of 20% Triton X-100, and H_2O to a final volume of 7 ml was added, followed by 1 hour of incubation at 37°C under constant shaking.

Ligation was performed by adding 70 μl of 100 mmol/l ATP (Roche) and 50 Weiss Units (WU) of DNA Ligase (Fermentas/ThermoFisher, Waltham, USA). The sample was incubated for 5 hours at 16°C. During incubation, additional 10 WU of DNA ligase were added. Following ligation, 30 μl 10 mg/ml proteinase K (Qbiogene; MP Biomedicals, Santa Ana, CA, USA) were added, and the

sample was incubated overnight at 65°C. Next morning, 30 μl of 10 mg/ml RNase A (Roche) were added, and the sample was incubated for 30 minutes at 37°C.

The DNA was purified by two chloroform:phenol extractions, followed by ethanol precipitation using 1 ml 3 mol/l sodium acetate, 7 ml H_2O and 25 μl glycogen, taken up to a final volume of 50 ml with ice-cold ethanol. The mixture was kept overnight at −80°C. The pellet was finally resuspended in 150 μl H_2O.

Pre-digestion control, post-digestion control, and the final 3C sample (120 ng of DNA each) were analyzed on 1.5% agarose gels. Samples with satisfactory digestion were then pooled to proceed further.

The 3C samples were digested with a final quantity of 0.2 U/μl of the secondary restriction enzymes *Dpn*II or *Nla*III, respectively (New England Biolabs). The 4C digested samples were analyzed on an agarose gel. For the 4C ligation, 700 μl of T4 Ligase Buffer (Fermentas/ThermoFisher), 70 μl 100 mmol/l ATP, and 50 WU of DNA Ligase (Fermentas/ThermoFisher), were taken up to 7 ml with H_2O; this mixture was added to the samples, and the ligation reaction was incubated for 5 hours at 16°C. Finally, the samples were purified by phenol:chloroform extraction, followed by ethanol precipitation, and stored at −20°C.

For each viewpoint, 16 PCRs (for detailed PCR conditions and primer sequences, see Additional file 17: Table S1) were set up, using 30 ng of 4C template for each reaction. For ease of later Illumina library preparation, primers of a subset of samples were designed with an Illumina sequencing adapter tail (batch 1: *MEA F6*, *MEA F8*, *PHE*, *FIS2*, *CKI1*, *FWA*, *AG*, *FLC*). For all other samples (batch 2: *AT1G51860*, *AT3G44380*, *SWN*, *hk4s*, *YAO*), Illumina sequencing adapters were ligated later in the library preparation process.

An aliquot of each PCR product was analyzed on an agarose gel, and the remaining PCR product was purified using the QIAquick PCR Purification Kit (Qiagen, Hilden, Netherlands), following the manufacturer's protocol.

Library preparation

Hereafter, library preparation is described for samples that had no Illumina (Illumina, San Diego, CA, USA) adapter attached to the 4C primer. Samples of each replicate were pooled in equimolar amounts, and assessed on a Bioanalyzer (Agilent Technologies, Santa Clara, CA USA). Finally, each sample volume was adjusted to 100 μl using H_2O. Replicates were then split into two aliquots of 50 μl each, and 10 μl of Resuspension Buffer (RSB; Illumina) and 40 μl End-Repair Mix (ERP) (Illumina) was added. The mixture was incubated for 30 minutes at 30°C. Then, 100 μl of Agencourt AMPure beads (Beckman Coulter, Brea, CA, USA) were added, and the mixture was incubated for 15 minutes at room temperature. The reaction tubes were then placed on a magnetic stand. The supernatants were removed without disturbing the beads, and 400 μl of freshly prepared 80% ethanol was added. After 30 seconds, the ethanol was replaced with another 400 μl of 80% ethanol. The supernatant was removed, and the tubes were left open to dry. The beads binding the 4C PCR products were resuspended in 17.5 μl RSB, and incubated for 2 minutes before being placed on a magnetic stand for 15 minutes. Finally, 15 μl of sample was transferred to a fresh 0.2 ml reaction tube. To each sample, 2.5 μl of RSB and 12.5 μl A-tailing Mix (ATL) (Illumina) were added and mixed thoroughly, followed by incubation at 37°C for 30 minutes. Following this, 2.5 μl of RSB, 2.5 μl of DNA Ligase Mix (LIG) (Illumina) and 2.5 μl of indexed DNA adapters (Illumina) were added, and mixed gently by pipetting the mixture up and down. Subsequently, the mixture was incubated for 10 minutes at 30°C. To inactivate the reaction 5 μl of Stop Ligase Mix (STL) (Illumina) were added, and samples were transferred to a fresh 1.5 ml reaction tube. Then 42.5 μl of Agencourt AMPure beads (Beckman Coulter) were added to each tube, and the mixture was incubated for 15 minutes at room temperature. The tubes were subsequently placed on a magnetic stand for 2 minutes, then 80 μl of supernatant were removed and replaced with 200 μl of freshly prepared 80% ethanol. After incubation for 30 seconds, the supernatant was removed, and the tubes were left open to dry. The previous ethanol washing step described above was repeated once, then, the pellet was resuspended in 52.5 μl RSB. After 2 minutes of incubation at room temperature, tubes were placed on a magnetic stand for 2 minutes, then 50 μl of the supernatant were transferred to a fresh 1.5 ml reaction tube. The Agencourt AMPure (Beckman Coulter) cleanup was repeated once; however, at the final step, instead of being suspended in 52.5 μl RSB, the pellet was resuspended in 22.5 μl RSB, of which 20 μl were transferred to a fresh 0.2 ml reaction tube. Samples with adapters already attached to the 4C PCR primers were treated in the same way from this point on. To perform final library amplification, 5 μl of PCR Primer Cocktail (PPC) and 25 μl of PCR Master Mix (PMM) (both Illumina) were added to each tube. PCR was performed under the following conditions: 98°C for 30 seconds; then 12 cycles of 98°C for 10 seconds, 60°C for 30 seconds, and 72°C for 30 seconds; followed by a final elongation at 72°C for 5 minutes. Samples were then transferred to a 1.5 ml reaction tube, and 50 ml of Agencourt AMPure beads (Beckman Coulter) were added. After 15 minutes of incubation at room temperature, the tubes were placed on a magnetic stand for 2 minutes. Following this, 95 μl of supernatant were removed, and the beads were washed twice with 200 μl of freshly prepared 80% ethanol. After the supernatant was removed, tubes were left open to dry. The pellet was then resuspended in 32.5 μl RSB and

incubated for 2 minutes at room temperature. The tubes were placed on a magnetic stand, and 30 μl of the purified library were transferred to a fresh 1.5 ml reaction tube. From each library a 10 nmol/l stock in Tris-Cl (pH 8.5) with 0.1% (v/v) Tween 20 was prepared. All replicates in the libraries were subsequently pooled, and used for Illumina HiSeq 100 bp single end sequencing. For each batch of replicates, one lane per replicate was loaded (total of four lanes). Batch 1 replicate A had a total yield of 92,063,669 raw reads, with a mean quality score of 35.35. Batch 1 replicate B had a total yield of 80,777,012 raw reads with a mean quality score of 35.31; batch 2 replicate A had a total yield of 43,296,252 raw reads with a mean quality score of 36.85; and batch 2 replicate B had a total yield of 55,187,969 raw reads with a mean quality score of 36.76.

4C sequencing data pre-processing

The two fastq files (one per replicate) were split into separate viewpoints according to the 4C primer sequences and the *Hind*III restriction pattern within the reads. No mismatches were allowed, and the remaining reads were discarded. After removal of primer and restriction site sequences, reads were trimmed to 30 bp and aligned to the *Arabidopsis* reference genome [48] using bowtie (version 0.12.7) [49] with the command line arguments -a -v 0 -m 25. For alignment statistics, see Additional file 17: Table S2.

Reads with multiple alignments were processed as described previously [50]. Because we estimated the length of a single interaction unit as 100 kb, we used an allocation distance of ±50 kb. To specify potential 4C fragments, we generated an *in silico Hind*III digest of the *Arabidopsis* Col-0 genome. Reads mapping to the ends of the resulting fragments were considered for further analysis. For a more robust measure of interactions, fragments were then used to generate windows spanning a larger region of the genome (that is, 100 fragments, corresponding to 180 kb on average). During this process, fragments closer than 1 kb to the viewpoint were discarded, given that a large proportion of their reads would probably originate from incomplete digestion and/or self-circularization. Furthermore, we discarded all fragments closer than 100 kb to a centromere, as the quality of alignments to centromeres is low. Finally, fragments whose distance from the primary restriction site to the first occurring secondary restriction site was 1000 bp or more with respect to both ends of the fragment were also removed. As a measure of interaction of a given window (interaction value), fragment counts were log-transformed to avoid high impact of outlier fragments, and then summed. Depending on the downstream analysis, windows spanned either 100 fragments from each fragment on (overlapping) or 25 fragments starting from every 25th fragment (non-overlapping).

Processed 4C data files (split according to primer sequence) and raw-data sequencing files are publically available on Gene Expression Omnibus (GEO), accession number GSE50181.

Data processing of histone modifications, transcription, DNA methylation, and genomic sequencing

To add additional information, such as histone modification patterns and transcription rates, we obtained publicly available data from GEO [51], specifically ChIP sequencing (ChIP-seq) data GSM701923, GSM701924, GSM701925, GSM701926, GSM701927, GSM701928, GSM701929, GSM 701930, GSM701931 [30], and RNA-seq data GSM701934 [30]. Pre-processed DNA methylation data was obtained from [32].

ChIP-seq and RNA sequencing (RNA-seq) reads (SOLiD sequencing, 50 bp (Applied Biosystems/Life Technologies) were aligned to the *Arabidopsis* reference genome (Col-0, TAIR10 [52]) using bowtie (version 0.12.7) with the following command line arguments: –a –v 2 –m 25. Reads with multiple alignments were processed as described previously [50]. Allocation distances were set to ±5 kb and ±50 bp for the ChIP-seq and the RNA-seq data, respectively. Histone modification densities and DNA methylation densities were calculated by the sum of nucleotides covered by at least one uniquely alignable short sequence, divided by the total number of nucleotides for each individual 4C restriction fragment.

To estimate potential biases related to sequence composition (such as repetitive sequences), we obtained genomic DNA sequencing data (Illumina, 100 bp) of the data set GSM567816, and processed them identically to the 4C sequencing data.

Assigning *P*-values to individual windows

To estimate the significance of an interaction, we calculated for each window the probability (that is, *P*-value) to observe its interaction value by chance. Given that an interaction of two fragments would lead to a higher read count in the neighboring fragments as well (hence in the window), random shuffling of fragment positions and recalculation of window interaction values provides randomized interaction data with the values following a normal distribution. Using the parameters of this distribution, a preliminary *P*-value was then calculated for each window. We repeated this process 1,000 times, and averaged for each window the *P*-values from all individual repetitions to obtain a final *P*-value. To take into account the differences between chromosome arms (for example, the different amount of DNA between the short arm and the long arm of chromosome 2), the *P*-values were calculated for each chromosome arm separately.

P-value thresholds were chosen to best fulfill the requirements of either plotting or data analysis. Generally, we set the threshold for prey regions to 10^{-3}. In the Circos plot of

Figure 5A we chose $P \leq 10^{-4}$ for better visibility. Because for various viewpoints, a threshold of 10^{-3} did not yield a sufficient number of prey regions for robust data analysis, we chose a threshold of $P \leq 0.05$ to perform PCA.

Distance decay

We estimated the genomic distance-dependent decay of the interaction probability on a distance of 1 kb to 10 Mb from the viewpoint. This stretch was log-transformed, and split into 41 intervals with length of 0.1 (on the log scale). For each sample, the reads of the fragments corresponding to the intervals were summed up and assigned to the interval. Given that the centromere acts as an interaction boundary, only fragments on the viewpoint's arm were considered. Read counts per interval were then divided by the total number of reads across all intervals representing contact probabilities, which across the full distance add up to 1. Given that some intervals contained only a few fragments and, in certain cases, only fragments from a subset of the viewpoints, we used a locally weighted scatterplot smoothing (LOESS) predictor fitted to the original data to calculate one single contact probability value for each interval. To obtain the slope, and hence the distance decay coefficient, we then approximated the data with a linear model. Slope and *P*-value were derived from the fit of the linear model to the values predicted by the LOESS fit. However, direct fitting of a linear model to the original data yielded almost equal results with a slope of −0.72 instead of −0.73, and an extremely low *P* value ($<10^{-100}$).

Centromere distance

To analyze the effect of a viewpoint's distance to the centromere on the distribution of the observed interaction frequencies along chromosome arms, we calculated for each chromosome arm (except the viewpoint's arm) the distance to the centromere at which 50% of all reads were aligned, and then fitted a linear model. The procedure was performed twice, first using absolute values, and then relative distances, defined as the absolute distance divided by the length of the chromosome arm (transformed by taking the arcsine of the square root).

Principal component analysis

All PCAs were based on non-overlapping windows that included 25 fragments. For each viewpoint, mean prey and control histone densities for each histone modification (that is, EMD) were calculated. Subsequently, PCA was performed on a dataset including mean EMD values of control and prey regions for each viewpoint and EMD. PCA was performed using the built-in R princomp() function.

Permutation test

To analyze differences in the epigenetic landscape of prey and control regions, we randomly selected 50 prey and 50 control regions (sampled) for each viewpoint, and obtained a corresponding randomized test set by pooling their EMDs and permuting them (shuffling them into two randomized groups of 50 values each). We then calculated the absolute differences in averaged EMDs between the sampled ($RealDiff_{ij}$), and the permutated ($RandDiff_{ij}$) prey and control regions, respectively.

Repeating this step *i* times for each of the *j* viewpoints yielded an empirical distribution for RandDiff for every epigenetic modification with 13,000 values (j = 13 viewpoints, and i = 1,000 repetitions). Comparing the average $RealDiff_m$ (mean across all repetitions and viewpoints) with this distribution then provided an empirical *P*-value ($p = \sum(RandDiff_{ij} > RealDiff_m)/(i*j)$), which was subsequently adjusted for multiple testing calculating false discovery rate (FDR; Benjamini-Hochberg).

Analysis of individual epigenetic marks employing GSEA-like analysis

To test whether prey regions have a different epigenetic landscape from that of regions chosen randomly across the genome, we developed a procedure similar to the GSEA described previously [33]. It requires densities of EMDs (for example, CG methylation density or H3K9me2) assigned to all (*n*) regions in the genome (that is, non-overlapping windows spanning 25 restriction fragments), and a subset (*m*) of the regions as a test set (that is, prey regions with a $P < 0.01$ in both replicates). During the procedure, the regions are first sorted according to their EMD. We then assigned a value of −1 to regions not in the test set, and a value of $(n\text{-}m)/m$ to the regions in the test set (to assure that the sum of these values across all regions would be zero). In a third step, the cumulative sum of these values was calculated and the enrichment score (ES) was defined as the maximum (absolute) deviation from zero. If the regions in the test set were randomly distributed across the sorted list of all regions, the cumulative sum would fluctuate around zero with a relatively small ES. Conversely, a non-random distribution of the test set (for example, accumulation at one end of the sorted list) would lead to a high ES. A *P*-value could then be assigned by comparing an observed ES to an ES distribution obtained by randomly choosing *m* regions 10,000 times. To obtain one *P*-value per epigenetic feature, the ES were averaged across all viewpoints. As we were focusing on long-range interactions, we excluded all interactions within the viewpoint's arm. Because statistical testing for all epigenetic features was employed, using the same 4C data, *P*-values were adjusted for multiple testing, calculating FDR (Benjamini-Hochberg).

Plotting

All plotting of 4C data, genomic features, and histone modification data was performed using either Circos

[23] or built-in R functions [53] plotting. Code is available upon request.

Data availability

All sequencing data and processed 4C files are available on Gene Expression Omnibus (GEO) accession number GSE50181.

Additional files

Additional file 1: Figure S1. Circular chromosome conformation capture (4C) interactome of *MEA* F6.

Additional file 2: Figure S2. Circular chromosome conformation capture (4C) interactome of *MEA* F8.

Additional file 3: Figure S3. Circular chromosome conformation capture (4C) interactome of *AT1G51860*.

Additional file 4: Figure S4. Circular chromosome conformation capture (4C) interactome of *PHE1*.

Additional file 5: Figure S5. Circular chromosome conformation capture (4C) interactome of *FIS2*.

Additional file 6: Figure S6. Circular chromosome conformation capture (4C) interactome of *CKI1*.

Additional file 7: Figure S7. Circular chromosome conformation capture (4C) interactome of *AT3G44380*.

Additional file 8: Figure S8. Circular chromosome conformation capture (4C) interactome of *SWN*.

Additional file 9: Figure S9. Circular chromosome conformation capture (4C) interactome of *hk4s*.

Additional file 10: Figure S10. Circular chromosome conformation capture (4C) interactome of *YAO*.

Additional file 11: Figure S11. Circular chromosome conformation capture (4C) interactome of *AG*.

Additional file 12: Figure S12. Circular chromosome conformation capture (4C) interactome of *FWA*.

Additional file 13: Figure S13. Circular chromosome conformation capture (4C) interactome of *FLC*.

Additional file 14: Figure S15. Principal component analysis (PCA) for individual viewpoints. Each graph represents a bi-plot of a PCA, including histone modification densities (EMDs) for prey and control regions of a given viewpoint, respectively. Contributions to the variance of the first two principal components are indicated below the bi-plot. Loadings of the four major factors to the first principal component are listed.

Additional file 15: Figure S16. Epigenetic modification density (EMD). For each EMD and viewpoint, the mean EMD for 1,000 × randomly chosen 50 prey and control regions was calculated and plotted. Green bars, prey; yellow bars, control.

Additional file 16: Figure S14. Interaction frequency decay for individual viewpoints. Interaction frequency decay is plotted for individual viewpoints. Black line: LOESS smoothened decay. Red dotted line: Linear regression. Values of the slopes are indicated in the lower left corner of each graph.

Additional file 17: Table S1. Viewpoint coordinates and primer sequences. Indicated are the viewpoints' names, their respective chromosome and position in bp, primer sequences, and restriction enzymes used for primary (1°RS) and secondary (2°RS) digest, respectively. **Table S2.** Alignment scores. Columns indicating chromosomes show numbers of mapped reads. Other columns show unmapped reads, percentage of mapped reads, and total reads.

Abbreviations

3C: Chromosome conformation capture; 4C: Circular chromosome conformation capture; ChIP-seq: Chromatin immunoprecipitation sequencing; EMD: Epigenetic modification density; ES: Enrichment score; FDR: False discovery rate; FISH: Fluorescent *in situ* hybridization; GEO: Gene Expression Omnibus; GSEA: Gene Set Enrichment Analysis; H3K27me1: Monomethylation of lysine 27 of H3; H3K36me3: Trimethylation of lysine 36 of H3; H3K4me2: Dimethylation of lysine 4 of H3; H3K9me2: Dimethylation of lysine 9 of H3; PCA: Principal component analysis; RNA-seq: RNA sequencing; RPKM: Reads per kilobase per million; RPM: Reads per million.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SG conceived the study, conducted the experiments, performed data analysis, and wrote the manuscript. MWS performed data analysis and helped to write the manuscript. TW helped to conceive the study and helped to edit the manuscript. NL helped to conceive the study. UG conceived the study, and helped with data interpretation and writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Keith Harshman, Johann Weber, and Corinne Peter (University of Lausanne) for advice on Illumina library construction, and Heike Lindner, Aurélien Boisson-Dernier, and Pauline Jullien for critically reading the manuscript. This work was supported by the University of Zürich, the University Research Priority Program Functional Genomics/Systems Biology, an IPhD project grant from SystemsXch, the Swiss Initiative for Systems Biology (to UG, TW, and NL), and an Advanced Grant of the European Research Council (to UG).

Author details

[1]Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, CH-8008 Zürich, Switzerland. [2]Institute of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland.

Received: 18 June 2013 Accepted: 24 November 2013
Published: 24 November 2013

References

1. Dittmer TA, Stacey NJ, Sugimoto-Shirasu K, Richards EJ: ***LITTLE NUCLEI* genes affecting nuclear morphology in *Arabidopsis thaliana*.** *Plant Cell* 2007, **19:**2793–2803.
2. Arnott S, Hukins DW: **Optimised parameters for A-DNA and B-DNA.** *Biochem Biophys Res Commun* 1972, **47:**1504–1509.
3. Roudier FCO, Ahmed I, Rard CBE, Sarazin A, Mary-Huard T, Cortijo S, Bouyer D, Caillieux E, Duvernois-Berthet E, Al-Shikhley L, Giraut LEN, s BDE, Drevensek SEP, Barneche FED, Rozier SDE, Brunaud VER, Aubourg SEB, Schnittger A, Bowler C, Martin-Magniette M-L, Robin SEP, Caboche M, Colot V: **Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*.** *EMBO J* 2011, **30:**1928–1938.
4. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B: **Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells.** *Cell* 2010, **143:**212–224.
5. Pfluger J, Wagner D: **Histone modifications and dynamic regulation of genome accessibility in plants.** *Curr Opin Plant Biol* 2007, **10:**645–652.
6. Rabl C: **Über die Zelltheilung.** *Morphologisches Jahrbuch* 1885, **10:**214–330.
7. Heitz E: **Das Heterochromatin der Moose.** *1 Jahrb Wiss Bot* 1929, **69:**762–818.
8. La Cour L: **Heterochromatin and the organization of nucleoli in plants.** *Heredity* 1951, **5:**37.
9. Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D: **The dynamic architecture of *Hox* gene clusters.** *Science* 2011, **334:**222–225.
10. Gheldof N, Smith EM, Tabuchi TM, Koch CM, Dunham I, Stamatoyannopoulos JA, Dekker J: **Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene.** *Nucleic Acids Res* 2010, **38:**4325–4336.
11. McClintock B: **Chromosome morphology in *Zea mays*.** *Science* 1929, **69:**629.

12. Fransz PF, Armstrong S, de Jong JH, Parnell LD, van Drunen C, Dean C, Zabel P, Bisseling T, Jones GH: **Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: structural organization of heterochromatic knob and centromere region.** *Cell* 2000, **100:**367–376.
13. Laboratory TCSH, Washington University Genome Sequencing Center, Consortium PBAS: **The complete sequence of a heterochromatic island from a higher eukaryote.** *Cell* 2000, **100:**377–386.
14. Fransz P, Armstrong S, Alonso-Blanco C, Fischer TC, Torres-Ruiz RA, Jones G: **Cytogenetics for the model system *Arabidopsis thaliana*.** *Plant J* 1998, **13:**867–876.
15. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation.** *Science* 2002, **295:**1306–1311.
16. De Wit E, De Laat W: **A decade of 3C technologies: insights into nuclear organization.** *Genes Dev* 2012, **26:**11–24.
17. Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, Pant V, Tiwari V, Kurukuti S, Ohlsson R: **Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions.** *Nat Genet* 2006, **38:**1341–1347.
18. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326:**289–293.
19. Louwers M, Bader R, Haring M, Van Driel R, De Laat W, Stam M: **Tissue- and expression level-specific chromatin looping at maize *b1* epialleles.** *Plant Cell* 2009, **21:**832–842.
20. Crevillen P, Sonmez C, Wu Z, Dean C: **A gene loop containing the floral repressor *FLC* is disrupted in the early phase of vernalization.** *EMBO J* 2012, **32:**140–148.
21. Moissiard G, Cokus SJ, Cary J, Feng S, Billi AC, Stroud H, Husmann D, Zhan Y, Lajoie BR, McCord RP, Hale CJ, Feng W, Michaels SD, Frand AR, Pellegrini M, Dekker J, Kim JK, Jacobsen S: **MORC family ATPases required for heterochromatin condensation and gene silencing.** *Science* 2012, **336:**1448–1451.
22. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the *Drosophila* genome.** *Cell* 2012, **148:**458–472.
23. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res* 2009, **19:**1639–1645.
24. Splinter E, de Wit E, van de Werken HJG, Klous P, De Laat W: **Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation.** *Methods* 2012, **58:**221–230.
25. Hou H, Zhou Z, Wang Y, Wang J, Kallgren SP, Kurchuk T, Miller EA, Chang F, Jia S: **Csi1 links centromeres to the nuclear envelope for centromere clustering.** *J Cell Biol* 2012, **199:**735–744.
26. de Nooijer S, Wellink J, Mulder B, Bisseling T: **Non-specific interactions are sufficient to explain the position of heterochromatic chromocenters and nucleoli in interphase nuclei.** *Nucleic Acids Res* 2009, **37:**3558–3568.
27. Fransz P, De Jong JH, Lysak M, Castiglione MR, Schubert I: **Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate.** *Proc Natl Acad Sci U S A* 2002, **99:**14584–14589.
28. Fang Y, Spector DL: **Centromere positioning and dynamics in living *Arabidopsis* plants.** *Mol Biol Cell* 2005, **16:**5710–5718.
29. Gheldof N, Tabuchi TM, Dekker J: **The active *FMR1* promoter is associated with a large domain of altered chromatin conformation with embedded local histone modifications.** *Proc Natl Acad Sci U S A* 2006, **103:**12463–12468.
30. Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, Nieuwland M, Simonis M, De Laat W, van Lohuizen M, van Steensel B: **Interactions among *Polycomb* domains are guided by chromosome architecture.** *PLoS Genet* 2011, **7:**e1001343.
31. Luo C, Sidote DJ, Zhang Y, Kerstetter RA, Michael TP, Lam E: **Integrative analysis of chromatin states in *Arabidopsis* identified potential regulatory mechanisms for natural antisense transcript production.** *Plant J* 2012, **73:**77–90.
32. Stroud H, Greenberg MVC, Feng S, Bernatavichute YV, Jacobsen SE: **Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome.** *Cell* 2013, **152:**352–364.
33. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545–15550.
34. La Bastide DM, Huang E, Spiegel L, Gnoj L, Tabata S, Kaneko T, Nakamura Y, Kotani H, Kato T, Asamizu E, Miyajima N, Sasamoto S, Kimura T, Hosouchi T, Kawashima K, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakayama S, Nakazaki N, Naruo K, Okumura S, Shinpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Sato S, *et al*: **Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**823–826.
35. Maul GG, Deaven L: **Quantitative determination of nuclear pore complexes in cycling cells with differing DNA content.** *J Cell Biol* 1977, **73:**748–760.
36. Schubert V, Berr A, Meister A: **Interphase chromatin organisation in *Arabidopsis* nuclei: constraints versus randomness.** *Chromosoma* 2012, **121:**369–387.
37. Jin QW, Fuchs J, Loidl J: **Centromere clustering is a major determinant of yeast interphase nuclear organization.** *J Cell Sci* 2000, **113:**1903–1912.
38. Duan Z, Andronescu M, Schutz K, Mcilwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS: **A three-dimensional model of the yeast genome.** *Nature* 2010, **465:**363–367.
39. Sanyal A, Baù D, Marti-Renom MA, Dekker J: **Chromatin globules: a common motif of higher order chromosome structure?** *Curr Opin Cell Biol* 2011, **23:**325–331.
40. Armstrong SJ, Franklin FC, Jones GH: **Nucleolus-associated telomere clustering and pairing precede meiotic chromosome synapsis in *Arabidopsis thaliana*.** *J Cell Sci* 2001, **114:**4207–4217.
41. Watanabe K, Pecinka A, Meister A, Schubert I, Lam E: **DNA hypomethylation reduces homologous pairing of inserted tandem repeat arrays in somatic nuclei of *Arabidopsis thaliana*.** *Plant J* 2005, **44:**531–540.
42. Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo D-H, Shi J, Gao Z, Han F, Lee H, Xu R, Allison J, Birchler JA, Jiang J, Dawe RK, Presting GG: **Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons.** *PLoS Genet* 2009, **5:**e1000743.
43. Nagaki K, Song J, Stupar RM, Parokonny AS, Yuan Q, Ouyang S, Liu J, Hsiao J, Jones KM, Dawe RK, Buell CR, Jiang J: **Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres.** *Genetics* 2003, **163:**759–770.
44. Henikoff S: **The centromere paradox: stable inheritance with rapidly evolving DNA.** *Science* 2001, **293:**1098–1102.
45. Berr A, Pecinka A, Meister A, Kreth G, Fuchs J, Blattner FR, Lysak MA, Schubert I: **Chromosome arrangement and nuclear architecture but not centromeric sequences are conserved between *Arabidopsis thaliana* and *Arabidopsis lyrata*.** *Plant J* 2006, **48:**771–783.
46. Nagaki K, Talbert PB, Zhong CX, Dawe RK, Henikoff S, Jiang J: **Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres.** *Genetics* 2003, **163:**1221–1225.
47. Shibata F: **Differential localization of the centromere-specific proteins in the major centromeric satellite of *Arabidopsis thaliana*.** *J Cell Sci* 2004, **117:**2963–2970.
48. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, Mueller LA, Bhattacharyya D, Bhaya D, Sobral BW, Beavis W, Meinke DW, Town CD, Somerville C, Rhee SY: **The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant.** *Nucleic Acids Res* 2001, **29:**102–105.
49. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10:**R25.
50. Schmid MW, Schmidt A, Klostermeier UC, Barann M, Rosenstiel P, Grossniklaus U: **A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing.** *PLoS ONE* 2012, **7:**e29685.
51. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30:**207–10.

52. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E: **The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools.** *Nucleic Acids Res* 2011, **40**:D1202–D1210.
53. Development Core Team R: *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008. http://www.R-project.org. ISBN 3-900051-07-0.

doi:10.1186/gb-2013-14-11-r129
Cite this article as: Grob *et al.*: Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture. *Genome Biology* 2013 **14**:R129.

8.2 Hi-C Analysis in *Arabidopsis* Identifies the *KNOT*, a Structure with Similarities to the *flamenco* Locus of *Drosophila*

The following manuscript is published in "Molecular Cell" (published by Elsevier Ltd, all rights reserved)[1]. I designed the data handling concepts and implemented the raw data preprocessing (see chapter 6). Stefan Grob and I designed the data analysis concepts together, and I implemented a part of it. I further contributed to data analysis and interpretation, wrote a part of the methods section, and helped to improve the manuscript.

[1] Grob, S, Schmid, MW, and Grossniklaus, U (2014) Hi-C Analysis in *Arabidopsis* Identifies the *KNOT*, a Structure with Similarities to the *flamenco* Locus of *Drosophila*. Molecular Cell 55: 678–693.

CellPress

Molecular Cell
Article

Hi-C Analysis in *Arabidopsis* Identifies the *KNOT*, a Structure with Similarities to the *flamenco* Locus of *Drosophila*

Stefan Grob,[1] Marc W. Schmid,[1] and Ueli Grossniklaus[1,*]
[1]Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, 8008 Zürich, Switzerland
*Correspondence: grossnik@botinst.uzh.ch
http://dx.doi.org/10.1016/j.molcel.2014.07.009

SUMMARY

Chromosomes are folded, spatially organized, and regulated by epigenetic marks. How chromosomal architecture is connected to the epigenome is not well understood. We show that chromosomal architecture of *Arabidopsis* is tightly linked to the epigenetic state. Furthermore, we show how physical constraints, such as nuclear size, correlate with the folding principles of chromatin. We also describe a nuclear structure, termed *KNOT*, in which genomic regions of all five *Arabidopsis* chromosomes interact. These *KNOT ENGAGED ELEMENT* (*KEE*) regions represent heterochromatic islands within euchromatin. Similar to PIWI-interacting RNA clusters, such as *flamenco* in *Drosophila*, *KEE*s represent preferred landing sites for transposable elements, which may be part of a transposon defense mechanism in the *Arabidopsis* nucleus.

INTRODUCTION

Eukaryotic nuclei represent highly complex structures and are involved in many cellular processes. The storage and reading of genetic information require elaborate packaging of chromosomes, which depends on two seemingly conflicting factors: condensation and accessibility of DNA.

Chromosomes are organized into distinct regions, referred to as chromosome territories (CTs). The two chromosome arms (CAs) of a CT form a tight interaction unit, clearly separated from each other (Grob et al., 2013; Schubert et al., 2012). In animals, CAs were initially subdivided into discrete chromatin domains that are distinguished by differential packaging densities and epigenetic state (Lieberman-Aiden et al., 2009). Less packaged domains are characterized by activating epigenetic marks, such as H3K4me3, whereas more densely packaged domains are enriched in the inactive epigenetic mark H3K27me3 (Sexton et al., 2012). Using higher resolution, our knowledge on mammalian chromatin organization could be refined by the finding of topological domains that are demarcated by an enrichment of the insulator protein CTCF (Dixon et al., 2012).

Interaction decay exponents (IDEs) describe the steepness of the slope with which chromatin interaction frequencies (IFs) obtained in Hi-C experiments decay with distance from a given viewpoint. IDEs were used to predict polymer-folding principles in human nuclei, for which distinct models, the fractal globule model (FGM) and the equilibrium globule model (EGM), were proposed (Lieberman-Aiden et al., 2009). The EGM suggests a densely packed polymer with various knots, in which different regions of the polymer interlace. The FGM describes a polymer structure that exhibits globular substructures, reminiscent of beads on a string. As the FGM lacks knots, allowing for easy untangling of chromosomes, it is convenient to describe chromatin conformation. Both models differ in their theoretical IDEs: FGM and EGM yield IDEs of −1 and −1.5, respectively. Several chromosome interaction studies reported IDEs supporting the FGM (Grob et al., 2013; Lieberman-Aiden et al., 2009; Sexton et al., 2012; Zhang et al., 2012). However, chromatin organization is unlikely uniform along a chromosome, being composed of constitutive heterochromatin in pericentromeric regions (PRs) and euchromatic CAs. Whether PRs and CAs exhibit different IDEs, reflecting a distinct chromatin organization, is not clear, but previous studies showed that IDEs can differ between chromatin states (Sexton et al., 2012).

In *Arabidopsis thaliana*, PRs and CAs clearly differ in appearance, with PRs being part of chromocenters, brightly DAPI-stained dots in interphase nuclei (Fransz et al., 2002). Thus, calculation of IDEs of different chromatin states promises more realistic insights into chromatin organization.

Nuclear architecture is expected to be influenced by extrinsic factors, including nuclear volume. CROWDED NUCLEI (CRWN1, CRWN2, CRWN3, and CRWN4) proteins control nuclear size and are localized to the nuclear periphery (Dittmer et al., 2007; Sakamoto and Takagi, 2013; Wang et al., 2013). In *crwn1* and *crwn4* mutants, nuclear size is up to 75% smaller. Additionally, *crwn4* mutants exhibit fewer and dispersed chromocenters, indicating a role in heterochromatin regulation. Although the effects of *crwn* mutants on nuclear morphology have been described, it remains unknown how these changes affect chromosomal architecture. Therefore, we analyzed chromosomal architecture by performing Hi-C experiments on nuclei of *crwn1* and *crwn4* mutant *Arabidopsis* seedlings.

To date, very few studies have been published assessing differences between wild-type (WT) and mutant Hi-C data sets. Thus, a gold standard on how to assess differences between Hi-C data sets is lacking. We propose a computational method to assess the significance of changes observed in different Hi-C data sets and report how *crwn1* and *crwn4* mutants affect

chromosomal architecture. Hi-C not only allows a description of the principles of chromatin organization but also identifies discrete chromosomal interactions, which might confer functional significance. We identified a structure consisting of an entanglement of ten chromosomal regions, the *KNOT*. As it shows certain similarities to the *flamenco* locus of *Drosophila*, which controls several transposable elements (TEs) by RNAi, we postulate a function of the *Arabidopsis* KNOT in TE regulation and processing.

RESULTS

To gain insight into the chromosomal architecture of *Arabidopsis* nuclei, we performed Hi-C experiments on WT, *crwn1-1*, and *crwn4-1* seedlings of the Columbia-0 (Col-0) accession.

Chromosomal Neighborhood

We sought to understand how CTs relate to each other and investigated the spatial distribution of chromosomes in the nucleus. We calculated the expected (Zhang et al., 2012) IFs for each pair of *trans*-interacting chromosomes and compared these values to the observed IFs between these pairs. The log-ratio between observed and expected Hi-C interactions was used to describe whether two given chromosomes interact more with each other than expected and hence are located in spatial proximity (Figure 2A). Deviations from expected IFs were low compared to a study in mice (Zhang et al., 2012), suggesting equal interactions between all five *Arabidopsis* chromosomes.

Hi-C Interactions Form Defined Interaction Domains

The relationship between interactions of neighboring genomic bins allows insight into chromosomal architecture. As previously shown (Lieberman-Aiden et al., 2009; Sexton et al., 2012; Zhang et al., 2012), Hi-C interaction values are not independent of each other but correlate, forming domains of interacting regions (Figures 1A and 1C). Two Hi-C bins in close genomic proximity should share common interactors as they are physically connected. To better define structural domains (SDs), we calculated correlation coefficients of the distance-normalized interaction matrix. Visualization of the distance-corrected correlation matrix facilitated the observation of distinct SDs (Figure 1B). The major domains of chromatin organization were limited to euchromatin of CAs and heterochromatin found in PRs (Table S1 available online and Figure 5C). Yet, we could detect additional SDs within euchromatic CAs encompassing several megabases (Figures 1B–1D and S1).

As previously reported (Grob et al., 2013; Moissiard et al., 2012), we observed increased IFs and high correlation between the PRs of the *Arabidopsis* chromosomes, indicating clustering within the nucleus. Likewise, telomeric regions were observed to specifically interact among each other. Interactions between telomeres and PRs were depleted, suggesting differential compartmentalization (Figures 1A and 1B). Generally, we observed low IFs between euchromatic CAs and PRs, further supporting our previous observation (Grob et al., 2013) that heterochromatin and euchromatin represent distinct interactomes within the nucleus.

Principal Component Analysis Reveals Distinct Chromatin States

By close inspection of the correlated Hi-C data, we observed discrete SDs, which appeared to highly interact among each other but exhibited rather low IFs with the rest of the genome. Thus, we termed them compacted structural domains (CSDs). In contrast, other SDs exhibited a loose state (loose structural domains [LSDs]), characterized by depleted IFs within them but enriched IFs with more distal regions both in *cis* and *trans*.

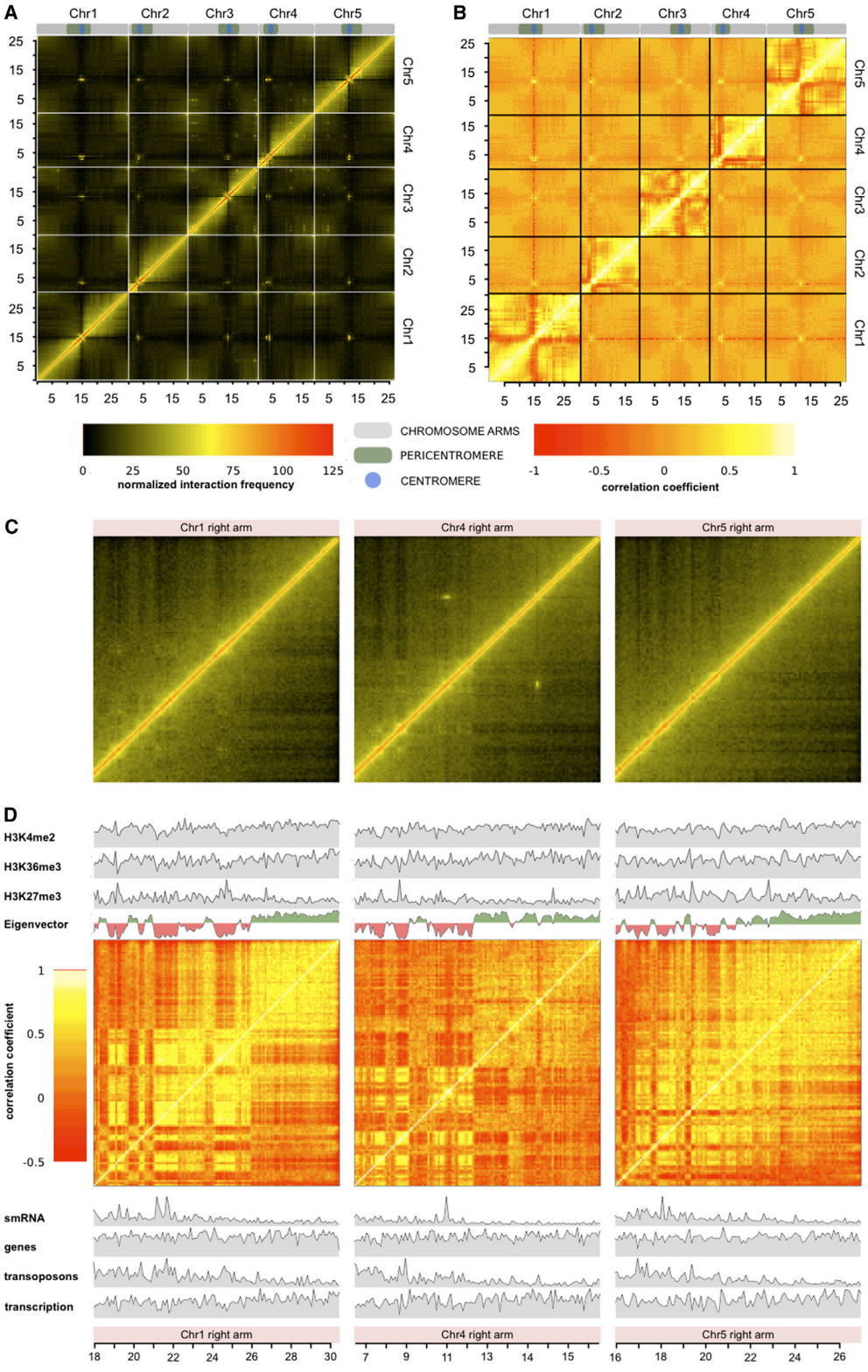
To obtain a numeric description of these SDs, we performed principal component analysis (PCA) on the correlation matrix of each individual chromosome (Chr). This led to a clear partitioning of the interactome into two SDs with either positive or negative Eigenvalues, with negative and positive Eigenvalues corresponding to CSDs and LSDs, respectively. The Eigenvalues can serve as a measure for domain structure, describing the accessibility—and therefore compaction state—of a given SD, and aid in accentuating the domain structure of chromatin (Figures 1C, 1D, and S1).

As expected, the first principal component (which describes the factor adding most to the variance of the data) was mainly dependent on the occurrence of constitutive heterochromatin or euchromatin, and it therefore hindered uncovering a detailed domain structure by PCA. To understand SD formation within euchromatin, we calculated correlations matrices and subsequently PCAs separately for each euchromatic CA, excluding heterochromatic PRs from analysis (Table S1). We found that the accentuation of discrete SDs varies between different CAs. The right arms of Chr1, Chr4, and Chr5 exhibited the clearest sequential arrangement of discrete SDs, whereas SDs on other CAs, although present, were less obvious (Figures 1B–1D and S1).

LSDs and CSDs Correlate with Epigenetic Chromatin States

Previous reports suggested a correlation between interactome and epigenome (Grob et al., 2013; Lieberman-Aiden et al., 2009; Sexton et al., 2012). Thus, we speculated that specific epigenetic marks correlate with LSDs and CSDs in CAs. To test this hypothesis, we obtained publicly available data on epigenetic and genomic features (see Supplemental Information). We computed Pearson's correlation coefficients between each feature and the Eigenvector for all euchromatic CAs individually (Figures 1D, 2B, and S2; Table S2). For the robustness of these analyses, the detection of discrete SDs is crucial. Therefore, we focused specifically on the right arms of Chr1, Chr4, and Chr5, which exhibited the most readily recognizable SDs (Figures 1D and S1).

Generally, histone modifications associated with active euchromatin (Filion et al., 2010; Roudier et al., 2011) exhibited strong correlations with the Eigenvector and highly significant p values. Specifically, high correlations were observed for H3K36me3 and H3K4me2, whereas strong anticorrelation was found for the *Polycomb*-associated mark H3K27me3 (Figures 2B and S2; Table S2). Histone marks associated with constitutive heterochromatin (H3K27me1, H3K9me2) showed weak anticorrelations. Of genomic features tested, transcription rate



(legend on next page)

highly correlated, whereas the number of TEs highly anticorrelated (Figures 2B and S2; Table S2). In summary, correlation analysis revealed that active histone modifications and transcription rate positively correlated with LSDs, whereas CSDs highly correlated with inactive epigenetic marks and genomic features of inactive euchromatin, such as abundance of TEs and accumulation of associated small RNAs (smRNAs).

To quantify the difference in epigenetic landscape between the two SDs, we assigned each genomic bin to one of two groups, defined by positive or negative Eigenvalues. To test whether the groups significantly differed in epigenetic landscape, we individually performed Wilcoxon rank sum tests for each feature and each CA (Figure 2C; Table S2). H3K9ac, H3K4me2, H3K4me3, H3K36me2, and H3K36me3 were significantly (α = 0.01) higher in LSDs for all CAs analyzed. The enrichment of active marks in LSDs varied little, with an average enrichment of 1.2- to 1.3-fold compared to CSDs over all CAs analyzed. In contrast, we observed significant enrichment of H3K27me3 in CSDs (1.3-fold) (Figure 2C).

Despite showing a significant enrichment in CSDs for a subset of CAs, density levels of H3K9me2 and H3K27me1 were generally low, further suggesting that histone modifications characteristic of constitutive heterochromatin do not play a major role for SD formation in euchromatic CAs. Although previously described to colocalize with H3K27me3 (Luo et al., 2012), we did not observe significant differences in H3K18ac (Figure 2C).

In plants, cytosine methylation occurs in the CG, CHG, and CHH context (where H is any base but G). In CSDs, DNA methylation in the CG, CHG, and CHH context was enriched 1.3-, 2.1-, and 1.8-fold, respectively. We observed a significantly higher transcription rate (1.5-fold) in LSDs, while gene density appeared to be a minor factor, as it was only negligibly higher in LSDs (1.1-fold). In contrast, the number of loci associated with smRNAs (2.1-fold) and TEs (2.4-fold) was significantly enriched in CSDs (Figure 2C). We could exclude that sequencing and alignment artifacts perturbed our analyses, as both the density of H3 occupancy and genomic sequencing reads did not significantly differ between LSDs and CSDs (Figure 2C). Furthermore, the results were robust using various genomic bin sizes (25, 50, and 100 kb).

In summary, we could detect a clear correlation between the spatial organization of chromatin and the epigenetic landscape. Features that are predominantly associated with epigenetically inactive euchromatin were enriched in CSDs, whereas features characteristic for active euchromatin were observed at higher densities in LSDs. As we excluded regions of known constitutive heterochromatin (e.g., PRs), we did not observe a correlation between epigenetic marks associated with heterochromatin and either LSDs or CSDs.

Arabidopsis Mutants Affecting Nuclear Size Affect the Interactome

We hypothesized that structural characteristics of nuclei could significantly influence chromosomal architecture. Nuclear size represents a likely factor affecting chromatin organization because it will limit the volume available to a CT. To investigate the effects of size constraints, we compared chromatin organization of WT nuclei with nuclei deficient for the structural components CRWN1 and CRWN4.

To investigate the impact of the *crwn1* and *crwn4* mutants, we calculated differences between all obtained Hi-C data according to a previously described method (Moissiard et al., 2012) (Figures 3A and S3). In short, we calculated the difference between all elements of two Hi-C matrices of interest. The resulting difference matrix was subsequently normalized according to the absolute IFs in the two Hi-C matrices of interest. By visual inspection of the plotted difference, we observed increased interchromosomal pericentromere interactions, increased interarm interactions, and slightly reduced intra-arm interactions in *crwn4* nuclei (Figures 3A and S3). The reduction of intra-arm interactions was most pronounced for interactions between PRs and more distal regions of the CAs. Complementarily, we observed increased interactions between the two halves of the PRs flanking the centromeres. In contrast, interactions within one-half of the PRs appeared to be depleted, and interactions of PRs and telomeres were reduced in *crwn4* nuclei.

Nuclei of *crwn1* showed similar changes in chromosomal architecture; however, differences to WT were less distinct and their overall magnitude was smaller (Figures 3A and S3). Generally, *crwn4* and *crwn1* nuclei exhibited enrichment in *trans*-interactions (both *trans*-arm and *trans*-chromosomal), suggesting higher genome-wide compaction in these mutants. These observations are consistent with previous studies (Dittmer et al., 2007; Sakamoto and Takagi, 2013), describing significantly smaller nuclei in *crwn* mutants, leading to space constraints and, thus, possibly higher *trans*-interactions among the chromosomes. Additionally, we observed increased IFs between the PRs of all five chromosomes (Figures 3 and S3).

Differences between *crwn1*, *crwn4*, and Col-0 Cluster in Defined Domains

As chromosomal architecture is partly influenced by stochastic factors, we expected that Hi-C data sets exhibit some variability not based on relevant biological differences. Therefore, we developed an analytical pipeline to reveal biologically significant changes between sets of Hi-C interactomes.

We made use of the axiom that regions in close genomic proximity, which are physically linked, correlate in their genome-wide interactomes. Thus, changes inflicted on the genome-wide interactome of a given genomic bin should be reflected by

Figure 1. Visualization of Hi-C Interactome

(A) Visualization of WT Hi-C IFs; genomic bin size: 250 kb.

(B) Visualization of distance-normalized WT correlation matrix; genomic bin size: 250 kb.

(C) Magnified view on right arms of Chr1, Chr4, and Chr5; bin size: 100 kb.

(D) Visualization of correlative interactomes of the CAs in (C). Eigenvector for each CA representing the Eigenvalues of each 100 kb genomic bin is shown. Additional tracks are densities of epigenetic modifications or number of genomic features.

See also Table S1 and Figure S1.

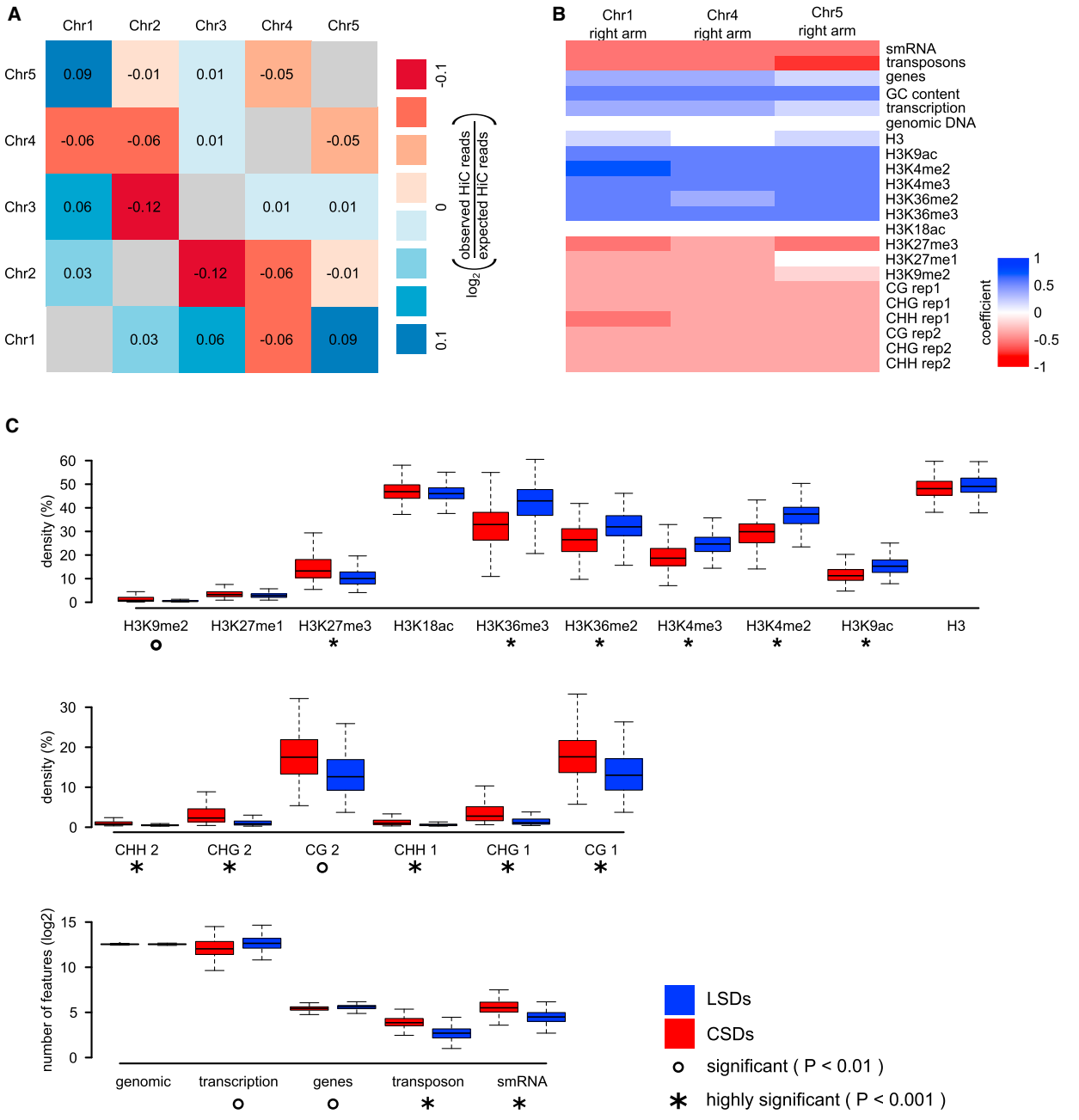


Figure 2. Chromosomal Neighborhood and Features Associated with Chromatin Organization

(A) Log_2 ratio of observed to expected pairwise interchromosomal interactions.

(B) Pearson's correlation coefficients between the Eigenvector (on 100 kb genomic bins) and epigenetic and genomic features for the right arms of Chr1, Chr4, and Chr5.

(C) Distribution of epigenetic and genomic features in LSDs and CSDs.

See also Table S2 and Figure S2.

changes in interactomes of neighboring genomic bins. We calculated matrix-wise correlation coefficients to obtain matrices of correlated differences (Figures 3B and S3). The representation of the correlation matrices showed that differences between Col-0 and the *crwn1* and *crwn4* mutants occurred in distinct domains.

To quantify this effect, we simplified the difference matrices, only considering whether a given interaction pair increases or decreases between two Hi-C data sets. This yielded a signed difference matrix (SDM) with the three possible elements: +, −, and 0 (for no difference) (Figures 3C, 3D, and S3). The Wald-Wolfowitz (WW) runs statistical test reveals whether the elements of a sequence are independent of each other. We expected that differences between two Hi-C data sets that arose from random noise in the data would be independent of each other for a given dimension of the matrix. Conversely, specific differences should occur in blocks of either positive or negative changes between the two Hi-C data sets. We calculated WW p values for each column in the SDM and counted the number of columns exhibiting a p value < 0.01; 50% of the genome-wide interactomes of genomic bins in the SDM of the pair *crwn4*-Col-0 exhibited significant p values. In comparison, 19% and 26% of the columns significantly differed in the *crwn1*-Col-0 and *crwn1-crwn4* SDMs, whereas only 2% significant differences were observed between two Col-0 replicates (Figure S3).

We then asked whether significant bins cluster along genomic positions. We expected significant columns to cluster if they contribute to changes that are based on biological differences between Hi-C data sets. Thus, we performed a second WW analysis, testing clustering of significant columns. This yielded extremely low p values for the pairs *crwn4*-Col-0, *crwn1*-Col-0, and *crwn1-crwn4*, but nonsignificant p values between two Col-0 replicates (Figures 3C and S3). In summary, alterations of chromosomal architecture associated with mutations in *crwn1* and *crwn4* clustered in defined domains, indicating a low contribution of stochastic variance to the observed differences.

SD Organization of CAs Does Not Change in *crwn1* and *crwn4* Mutants

Mutations affecting structural components of *Arabidopsis* nuclei influence *trans*-interactions. Intuitively, such alterations were expected due to the reduced nuclear size of *crwn1* and *crwn4* mutants, but they could also affect organizational differences within mutant nuclei. To study *cis*-interactions, and thus potential changes in local domain structure, we analyzed single chromosomes in more detail. We applied the above-described strategy to reveal SDs. As for WT nuclei, we focused our analysis on the right arms of Chr1, Chr4, and Chr5.

Making use of the Eigenvectors of each CA, we sought to detect potential changes in domain organization between WT and mutant nuclei. We individually performed cross-wise Pearson's correlation analyses between the different Hi-C data sets for all the three CAs (Figure 3E). Despite the observed alterations in *trans*-interaction patterns for a subset of mutants, we did not detect significant changes in the domain organization of CAs. The domain structure of all genotypes analyzed highly correlated among each other with negligible p values (Figure 3F). Consistent with this observation, we did not detect significant changes in SD organization when performing WW tests on the three CAs. As the only exception, we observed a minor change on the right arm of Chr1 when comparing *crwn1* to both WT and *crwn4*. We found an accentuated boundary between two SDs; this boundary encompassed the *CRWN1* gene and, in the *crwn1-1* mutant, the transfer DNA (T-DNA) insertion that caused the mutation (Figure 3F).

Hence, the SD organization of CAs appears to be a robust hallmark of chromosomal architecture, which is not significantly altered by mutations that affect nuclear size.

Distance-Dependent Decay of Interactions

Using distance-dependent mean interaction values, we can describe how IFs are coupled to the genomic distance of a given interaction pair. Previously, the distance-dependent decay of interactions, measured by IDEs, has been used to characterize chromatin packaging, specifically whether chromatin organization follows the EGM or FGM (Lieberman-Aiden et al., 2009).

IFs were shown to decay in a power-law function with an exponent of −0.867 (Figure 4A), consistent with previously described IDEs in *Arabidopsis* (Grob et al., 2013) and other organisms (Lieberman-Aiden et al., 2009; Sexton et al., 2012; Zhang et al., 2012). The variation of single chromosome IDEs was low, suggesting that all chromosomes share a common organization. To analyze how the IDE relates to different chromatin states, we calculated IDEs separately for PRs and for CAs (Figures 4B and 4C). Whereas variation within CAs and PRs was small (sd_{CA} = 0.02, sd_{PR} = 0.07), we noticed clear differences in IDE values between them. The mean IDE of PRs was −1.243 (Figure 4C), whereas CAs exhibited a smaller mean IDE of −0.704 (Figure 4B). Different IDEs of heterochromatic and euchromatic regions indicate a fundamentally different chromatin organization.

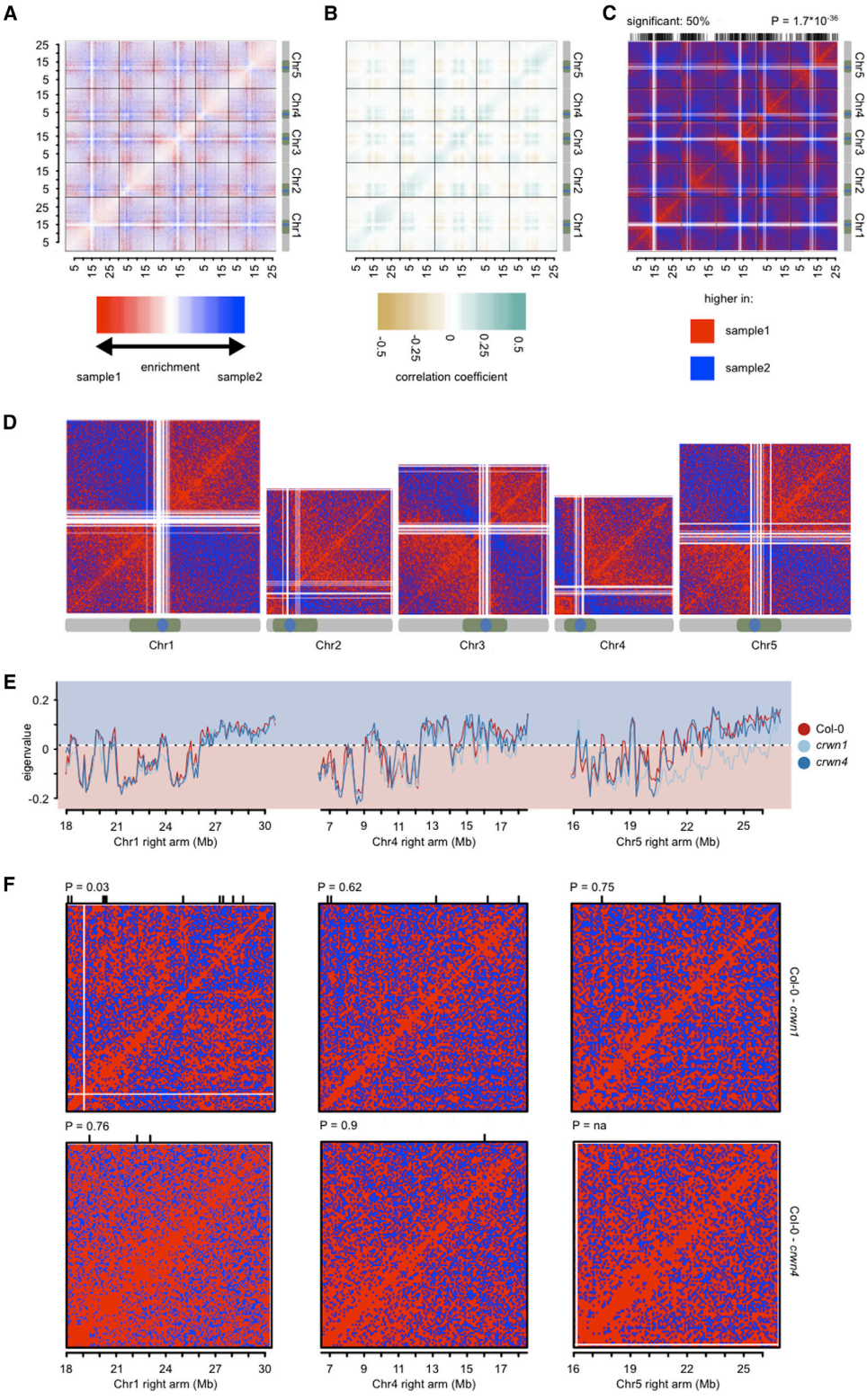
To reveal whether mutations affecting nuclear morphology such as *crwn1* and *crwn4* affect overall chromatin organization, we determined their genome-wide IDEs (IDE_{crwn1} = −0.834, IDE_{crwn4} = −0.846). These values are in agreement with the FGM of chromatin organization (Figure 4D). IDEs of PRs, however, exhibited clear differences between WT and mutant nuclei, implying differences in chromatin packaging. Pericentromeric IDEs of *crwn1* and *crwn4* were significantly higher than those of the WT (IDE_{crwn1} = −1.09, IDE_{crwn4} = −1.02; t test, p_{crwn1} = 0.006, p_{crwn4} = 0.001). This suggests an FGM of chromatin organization in PRs of mutant nuclei (Figure 4D).

In summary, Hi-C data sets differed considerably when their IDEs were calculated separately for PRs and CAs, indicating distinct packaging of these chromatin domains.

Specific Chromosome Interactions Form the *KNOT*

Visualizing raw Hi-C data, we observed discrete dots, likely representing highly specific interactions (Figures 1A and 5A). These dots seemed to connect a unique set of ten genomic regions, which appeared to interact almost exclusively among each other with high frequency (Figures 1A–1C and 5A). We concluded that all these genomic regions form an interacting structure that, in reminiscence of the nondisentangleable Gordian Knot (Plutarch, 1727), we termed the *KNOT*. The *KNOT* consists of both long- and short-range intrachromosomal as well as interchromosomal interactions. We found regions involved in the *KNOT* to reside on all chromosomes and named them *KNOT ENGAGED ELEMENT1* (*KEE1*) to *KEE10* (Figures 5B and 5C).

To unravel the nature of the ten *KEE*s, we identified their exact genomic position. We visualized each interaction pair of the *KNOT* separately at high resolution and estimated the genomic



(legend on next page)

coordinates of regions comprising the high-frequency interaction. As we expected a selected *KEE* to interact with all other *KEE*s with a defined core region, we hypothesized that this core should be reflected by the overlap of all pairwise interactions of the other *KEE*s with the selected *KEE*. Thus, we calculated the minimal overlap of all highly interacting regions for each *KEE*. With only one exception, all estimated core *KEE* positions overlapped each other (Figures 5B and S4), indicating that all *KEE*s interact within the *KNOT* with the same core position.

Fluorescence In Situ Hybridization Confirms the Existence of the *KNOT*

To independently confirm the robustness of the Hi-C data and the existence of the *KNOT*, we performed fluorescence in situ hybridization (FISH) on *Arabidopsis* seedling nuclei. We hybridized bacterial artificial chromosomes (BACs) to the chromatin of fixed leaf nuclei (Table S3). We selected BACs either encompassing *KEE*s or randomly chosen control regions. In each FISH experiment, we chose two distinctly labeled BACs in different combinations. These yielded nuclei in which either two *KEE*s, one *KEE* and one random region, or two random regions were labeled with different fluorescent markers (Figure 5D). Subsequently, association events between the two differentially labeled regions were analyzed by microscopy (Table 1; Figure 5F). As expected, we observed the highest association rates between regions located on the same chromosome, irrespective of whether the BACs encompassed *KEE*s or random regions.

However, we generally observed higher association rates between *KEE*s than between random regions. Strikingly, even *KEE*s separated by 20 Mb on different CAs showed higher association rates than a *KEE* and a random region located on the same CA and separated by only 6.1 Mb (Figure 5F). To analyze how the observed association rates relate to Hi-C interaction data, we performed in silico chromosome conformation capture (3C) experiments by calculating the sum of interactions between two regions (Figure 5E). Subsequently, by comparison of the Hi-C interaction values with the FISH association rates, we found the same interactions ranking high or low, respectively, in in silico 3C and FISH experiments (Figures 5E and 5F).

In summary, we could confirm the high IFs among *KEE*s by FISH and found comparable interaction and association rates, respectively, between FISH and Hi-C data.

*KEE*s Share Common Sequence Motifs

To better understand specific interactions among *KEE*s, we searched for common characteristics, such as sequence similarity. We extracted regions with high similarity using cross-wise alignments generated by the BLAST-like alignment tool (BLAT) (Kent, 2002), and we then refined the analysis with the motif search tool MEME (Bailey and Elkan, 1994). The highest similarity was detected for *KEE3*, *KEE4*, *KEE6*, *KEE7*, and *KEE9*, for which two motifs of 195 bp (motif1) and of 70 bp (motif2) were found (Figure S4).

To identify the genomic position of theses motifs, we performed BLAST searches and found that motif1 corresponded to TEs of the *ATLANTYS3* (LTR/Gypsy superfamiliy) and motif2 to*VANDAL6* (DNA MutR superfamiliy) families. Although not identified in the MEME search, we found *KEE2* and *KEE5* to exhibit significant sequence similarity with one of the two motifs. For the remaining *KEE*s, searching the genome with the sequence obtained in the BLAT-alignment, we found *ATLANTYS2* and a *TNAT1A* family DNA transposon (*KEE1*), *ATREP3*, *ATREP2*, and *VANDAL8* (*KEE8*), and *ATLANTYS3* and *VANDAL6* (*KEE10*).

In addition to the *KEE*s, we detected several other genomic regions that share sequence similarity with the motifs. These regions harbored *ATLANTYS3* and *VANDAL6* (Figure S4). We tested for increased IFs between these regions sharing sequence similarity with the *KEE*s. While *KEE*s exhibited significantly higher IFs among each other than with randomly chosen genomic bins ($p = 0.0004$), no enrichment of IFs was observed among regions sharing sequence homology to *KEE*s ($p = 0.2931$).

In summary, *KEE*s exhibit high sequence similarity, mainly corresponding to *ATLANTYS3* and *VANDAL6*. However, sequence similarity among *KEE*s is unlikely the crucial factor for formation of the *KNOT* because other genomic regions with sequence similarity to *KEE*s showed similar TE compositions but did not interact at high frequency.

*KEE*s Show a Specific Enrichment of Epigenetic and Genomic Features

As shown in this study, epigenetic features correlate with the interaction potential of a given region. To reveal common features, we analyzed the epigenetic landscape of *KEE*s (Figures 6A and 6B; Table S4). We observed a significant 2.7-fold enrichment of smRNAs associated with genomic regions surrounding the *KEE*s ($p = 0.0022$). For all other tested epigenetic and genomic features, we could not detect a significant enrichment or depletion in *KEE* regions ($\alpha = 0.05$; minimal enrichment or depletion: 1.5-fold).

We hypothesize that *KEE*s are not epigenetically homogenous as they are located in both PRs and CAs. If a genomic or epigenetic feature is characteristic for all *KEE*s, we postulate that the variance in density of that feature would be lower among *KEE*s than among randomly selected regions. However,

Figure 3. Comparison of WT to *crwn* Mutants

(A) Enrichment of IFs obtained by calculating the relative difference between Col-0 and *crwn4*.
(B) Pearson's correlation coefficients of differences between Col-0 and *crwn4*.
(C) Visualization of the SDM between Col-0 and *crwn4*.
(D) SDMs between Col-0 and *crwn4* for individual chromosomes.
(E) Comparison of the Eigenvectors of the right arms of Chr1, Chr4, and Chr5.
(F) Visualization of SDMs of individual CAs.
The lines on top of the SDM plots (C, E) indicate the location of genomic bins exhibiting significant ($\alpha < 0.01$) clustering of either positive of negative changes. (A)–(F) Genomic bin size: 100 kb.
See also Figure S3.

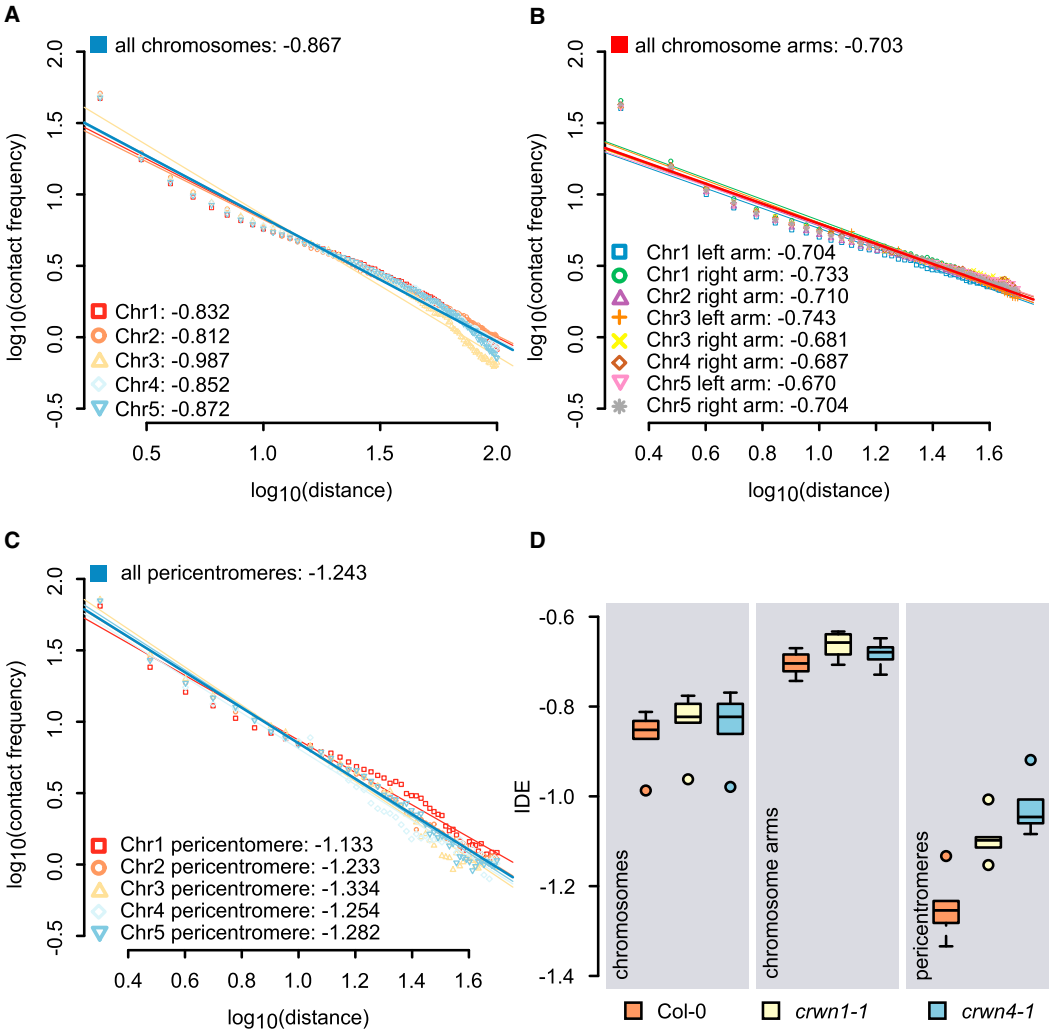


Figure 4. IDEs

(A) IDEs along chromosomes.

(B) IDEs along CAs.

(C) IDEs along PRs.

(D) Distribution of IDEs of the full genomes, CAs, and PRs for WT, *crwn1*, and *crwn4*.

In (A)–(C) dots represent average IFs between two regions of a given distance. The lines represent the fit of a linear model.

none of the investigated features varied significantly ($\alpha = 0.05$) less than expected. Consequently, we refined the analysis by considering only euchromatic *KEE*s (*KEE1*, *KEE3*, *KEE4*, *KEE6*, *KEE7*, *KEE8*, and *KEE9*) to reveal significantly enriched features. As in the above-described analysis for all *KEE*s, we found that smRNAs associated with *KEE* regions of 50 kb exhibited a significant 3.5-fold enrichment ($p < 0.0001$). In line with the observed enrichment of *VANDAL6* and *ATLANTYS3*, TEs were found two times more often in euchromatic *KEE*s than expected ($p = 0.0033$). Additionally, the heterochromatic mark H3K27me1 was 1.9-fold enriched ($p = 0.0119$) (Figures 6A and 6B; Table S4).

To confirm the robustness of these results, we repeated the analysis by testing for enrichment of a given feature within *KEE* regions of various size, i.e., 20, 50, 100, 150, 200, and 300 kb (Table S4). Whereas significant enrichments of smRNAs and H3K27me1 were observed in all window sizes tested, the enrichment of TEs was only significant in *KEE* regions of 50 and 100 kb. However, we additionally observed significantly increased transcription rates in *KEE*s, considering windows of 150, 200, and 300 kb.

Although rather heterogeneous concerning their epigenetic landscape, we conclude that *KEE*s in euchromatic CAs represent heterochromatic islands characterized by abundant

TEs, robust enrichment of smRNAs, and elevated levels of H3K27me1.

*KEE*s Are Preferred TE Insertions Sites

The occurrence of TEs, as well as the enrichment of smRNAs, led to the question whether *KEE*s play a role in TE processing, e.g., *KEE*s may represent a preferred TE landing site. A large number of insertion lines, consisting of several thousand independent events, are available in *Arabidopsis*. The majority of these lines were generated by *Agrobacterium*-mediated insertion of T-DNAs (SALK [Alonso et al., 2003], SAIL [Sessions et al., 2002], GABI-Kat [Kleinboelting et al., 2012], and FLAG [Samson et al., 2002]). Insertion lines created at Cold Spring Harbor Laboratory (CSHL) (Sundaresan et al., 1995) and RIKEN (Kuromori et al., 2004) were generated by the activation of a *Dissociation* (*Ds*) transposon and represent a collection of individual TE insertions. Wisconsin *Ds*Lox T-DNA lines (WISC) (Woody et al., 2007) were generated by *Agrobacterium*-mediated T-DNA insertion, but the vector also contained a *Ds* element.

We gathered information about the insertion sites of all available insertion lines from the SiGNAL database and tested for a preferential insertion into *KEE*s. We compared insertion frequencies within *KEE*s with insertion frequencies of 10,000 random sets of genomic regions. From the seven tested insertion collections, the two *Ds* populations (CSHL, RIKEN) exhibited a significant enrichment of insertions within *KEE*s (p_{CSHL} = 0.0003, p_{RIKEN} = 0.0008) (Figure 6D). All other analyzed collections, which were generated by T-DNA transformation (SALK, SAIL, GABI, FLAG, WISC), did not show significantly enriched insertion frequencies (Table S4). We also analyzed insertion sites of the retrotransposon *EVADÉ* (Marí-Ordóñez et al., 2013), which was reactivated in backgrounds with reduced DNA methylation (Mirouze et al., 2009). From 11 new *EVADÉ* insertions, 4 inserted within 250 kb of a *KEE* (Figure 6D).

In *Drosophila*, several PIWI-interacting RNA (piRNA) clusters are involved in TE silencing (Brennecke et al., 2007; Malone et al., 2009). Thus, we asked whether *Drosophila* piRNA clusters exhibit a similar interaction pattern as *KEE*s in *Arabidopsis*. Indeed, by inspection of previously published *Drosophila* Hi-C data (Sexton et al., 2012), we found significantly ($p < 0.0001$) enriched IFs between genomic regions harboring piRNA clusters (Brennecke et al., 2007) (Figure 6C).

In summary, we show that a *KNOT*-like structure is also formed by piRNA clusters in *Drosophila* and that *KEE*s are preferential insertion sites for TEs, suggesting a role in TE biology and thus genome integrity.

DISCUSSION

There Is No Distinct Chromosomal Neighborhood for a Given Chromosome

By calculating the deviation from the expected *trans*-IF between chromosomes, nuclear neighborhoods of CTs can be determined (Zhang et al., 2012). Compared to a study in mice (Zhang et al., 2012), the deviations from expected IFs in *Arabidopsis* nuclei are rather small. This suggests that any *Arabidopsis* chromosome has the same likelihood to stay in physical contact with any other, and that there is no preferential chromosome association. This conclusion is in line with FISH studies showing that *Arabidopsis* chromosomes do not preferentially pair (Pecinka et al., 2004).

The small number of chromosomes in *Arabidopsis* can explain the absence of distinct chromosomal neighborhoods. The higher number of chromosomes in mouse nuclei increases the probability that a chromosome is located between another pair, thereby separating distinct CTs. Single-cell Hi-C suggested a discrete number of interchromosomal contacts in a single mouse nucleus (Nagano et al., 2013). However, these contacts vary between nuclei of the same cell type, which leads to a rather uniform distribution of interchromosomal contacts in ensemble Hi-C, indicating that the preference of interchromosomal interactions is stochastic.

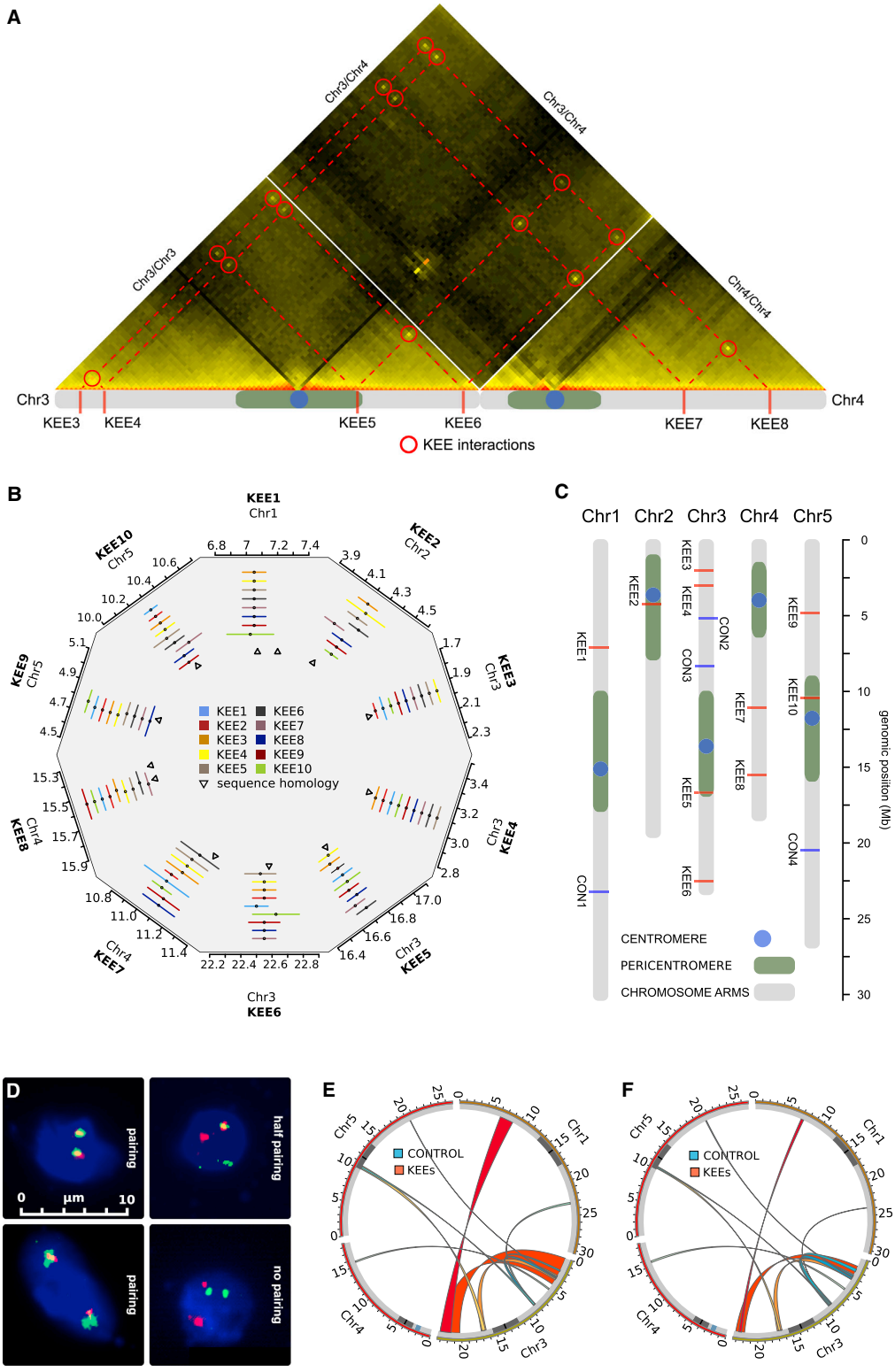
Arabidopsis Chromosomes Show a Simple Organization with Respect to Their Epigenetic Landscape and Interactome

Our results show that the epigenetic landscape strongly correlates with chromosomal architecture. LSDs, characterized by low compaction and enriched IFs with more distal regions both in *cis* and *trans*, are associated with active epigenetic marks, whereas the more condensed CSDs are enriched in repressive epigenetic marks. The composition of these two epigenetic landscapes is reminiscent of active chromatin state (CS) 1 and repressive CS2 (Roudier et al., 2011).

LSDs and CSDs resemble A and B compartments described in human cells. Similar to LSDs and CSDs, regions of the A compartment are less densely packaged than genomic regions of the B compartment (Lieberman-Aiden et al., 2009). The two classes of SDs in our study were distinguished by their inherent interaction potential. Thus, a single LSD or CSD can be subdivided into consecutive SDs with a similar interaction potential. These subdomains could be compared to topologically associating domains (Bickmore and van Steensel, 2013). However, we generally observed SDs to alternate in LSDs or CSDs, which seem to act as boundaries for each other.

Arabidopsis chromosomes show a rather simple organization with regard to the occurrence of constitutive heterochromatin and euchromatin. In all chromosomes, except Chr4, constitutive heterochromatin is solely found within PRs, whereas euchromatin is associated with CAs. The only additional region of constitutive heterochromatin of significant size, the knob *hk4s*, is on the short arm of Chr4 (Fransz et al., 2000; Grob et al., 2013). The organization of CAs is surprisingly homogenous, as all CAs exhibit increasing active marks, and therefore increasing occurrence of LSDs, toward distal positions. This makes it difficult to distinguish distinct SDs for a number of CAs.

The simple chromatin organization in *Arabidopsis* contrasts that of mammalian nuclei, in which CAs are clearly divided into numerous consecutive SDs (Lieberman-Aiden et al., 2009; Zhang et al., 2012). However, *Drosophila* nuclei exhibit a rather simple chromatin organization similar to that of *Arabidopsis* (Sexton et al., 2012). As the most conspicuous difference between mammalian genomes and those of *Drosophila* and *Arabidopsis* is their size, we propose that the highly compact nature of these genomes explains the apparent absence of structurally complex CAs.



(legend on next page)

Table 1. FISH Association Rates and Hi-C Interaction Scores

Probe 1	Probe 2	FISH Association Rate (%)	Hi-C Interaction Score
KEE6	*KEE1*	20	87.43
CON3[a]	CON1	3	5.44
CON3	*KEE3*	21	7.65
CON3	*KEE4*	31	8.36
KEE5	*KEE4*	35	34.96
KEE6	*KEE3*	66	92.39
KEE8	CON2	12	5.8
KEE5	*KEE10*	16	18.61
CON3	*KEE10*	9	4.11
CON4	*KEE4*	7	2

See also Table S3.
[a]CON, control BAC.

Nuclear Morphology Affects *trans*-Chromosomal Interactions but Not Domain Structure in *Arabidopsis* Nuclei

CRWN proteins are important structural components of *Arabidopsis* nuclei, with *crwn1* and *crwn4* mutants affecting nuclear morphology (Dittmer et al., 2007; Sakamoto and Takagi, 2013). *crwn1* and *crwn4* nuclei exhibited increased *trans*-interactions compared to WT nuclei, suggesting higher chromosomal compaction. As the size of *crwn1* and *crwn4* nuclei is substantially smaller than that of WT, we suggest that increased *trans* IFs are the consequence of size constraints, within which CTs have to be organized.

As a hallmark of chromosomal architecture in *crwn4* and, to a lesser extent, in *crwn1* nuclei, we observed increased interactions between PRs. Similarly, a reduced number of chromocenters and a disruption of chromocenter organization were cytogenetically observed in *crwn4* mutants (Wang et al., 2013). We conclude that this reduced number of observable chromocenters does not depend on chromatin decondensation but relates to an increased frequency of PR pairing, leading to the merging of PR territories.

The increased nuclear compaction in *crwn4* and *crwn1* nuclei is most obvious in the general increase of *trans*-arm interactions. In contrast, local chromatin organization within individual CAs is not grossly affected. This is evident by the unchanged occurrence of CSDs and LSDs within individual CAs. We conclude that chromosomes are organized in a hierarchical manner, in which CAs appear to be a stable unit, largely unaffected by physical constraints of nuclear morphology. However, CTs appear to be influenced by nuclear morphology. With less space available, two CA territories are forced into closer spatial proximity. Last, contacts between individual chromosomes appear to vary with nuclear size.

Variability in nuclear size and morphology is surprisingly high in *Arabidopsis*, which should influence *trans*-chromosomal interactions. However, much of this variation cannot easily be related to the transcriptional state of cells. Our results could provide a possible explanation for the lack of this relationship. The epigenetic landscape, and thus the transcriptional state of a cell, is mainly associated with the occurrence of SDs within CAs, which were shown to be largely independent of nuclear morphology.

Stochastic Variability between Interactomes Has to Be Carefully Assessed to Draw Biologically Relevant Conclusions

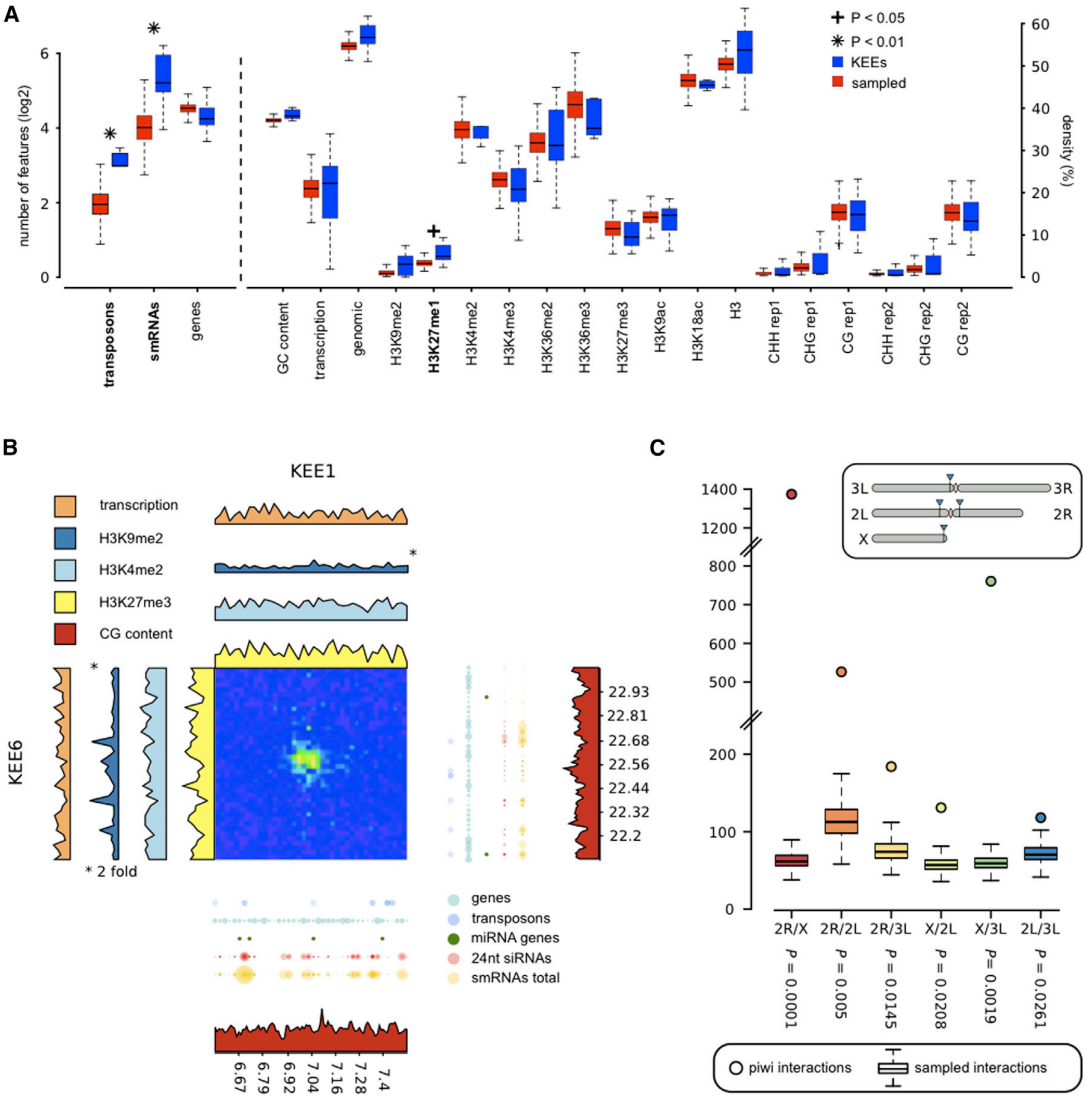
Chromosomal architecture is prone to stochastic variation, which is unlikely caused by important biological processes (Nagano et al., 2013). Therefore, careful assessment of this variability is essential for a conclusive evaluation of comparisons between different Hi-C data sets. We suggest an analytical pipeline to quantify stochastic variability, making use of the axiom that neighboring genomic bins should exhibit correlative interaction profiles.

The inspection of plain difference matrices bears the risk of overestimating the observation of patterns within these matrices. Hi-C interaction matrices are often visualized in symmetrical plots that represent a mirror image of the actual interactome, representing each interaction twice. This visualization method pronounces apparent patterns within the matrix, which would probably not be perceived as a distinct structure in a non-symmetrical visualization of the matrix. Analyzing correlative differences between two given Hi-C interaction data sets aids in a better understanding of the biological relevance of changes in Hi-C interactomes. Even more powerful, as it allows a statistical investigation of changes, is the analysis of whether clustered changes occur in SDMs, providing an even deeper insight into alterations of chromatin organization between different Hi-C data sets. As a major advantage, this method not only reveals genomic locations that undergo significant changes, but also provides an overall estimate of the difference between two interactomes by the total number of significant changes observed between them.

IDEs Indicate a Distinct Chromatin Organization of CAs and PRs

Most reported IDEs are close to the theoretical IDE of the FGM (*Drosophila*, −0.85 [Sexton et al., 2012]; mouse, −1.03 [Zhang et al., 2012]; human, −1.08 [Lieberman-Aiden et al., 2009]),

Figure 5. Positioning of *KEE*s, Shared Sequence Motifs, and FISH Validation
(A) Close-up of the interactions between and within Chr3 and Chr4. Red circles indicate high-frequency interactions between *KEE* regions.
(B) Estimated genomic intervals with the highest IF between a given *KEE* and all other *KEE*s (lines) and genomic positions of sequence homology among *KEE*s (triangles).
(C) Overview of the genomic positions of the *KEE*s on the five *Arabidopsis* chromosomes.
(D) Examples of FISH-analyzed nuclei. BACs are stained red and green, whereas DNA is stained blue.
(E) Circos plot of a virtual 3C experiment between *KEE* and control regions.
(F) Circos plot of FISH association rates.
(E and F) Red, interactions between *KEE*s; blue, interactions between control regions and between control regions and *KEE*s.
See also Figure S4.



(legend on next page)

indicating that the fractal globule organization is a conserved hallmark. The genome-wide IDE calculated in the present study (−0.895) further supports the FGM. By averaging IDEs of several circularized chromosome conformation capture (4C) experiments in *Arabidopsis*, we calculated an IDE of −0.73 (Grob et al., 2013). This value differs considerably from the genome-wide IDE calculated in the present study. However, in our previous work, 4C viewpoints were exclusively chosen within CAs. When we compared the IDE obtained by 4C experiments with the mean IDE of CAs in the present Hi-C study, we observed only a small difference between the two values (−0.73 and −0.703).

Interestingly, IDEs of different chromatin states differed considerably. Whereas euchromatic CAs exhibited an IDE of −0.703, the average IDE of PRs was −1.243. The IDEs of PRs suggest a different chromatin organization, which more closely resembles the EGM. This is not surprising as heterochromatin can easily be distinguished from euchromatin by its appearance. Therefore, accessibility, which is facilitated in a fractal globule chromatin organization, may not be a feature of heterochromatin. A different chromatin organization, such as the equilibrium globule organization, could be favorable to fulfill the requirements for heterochromatin.

Previous observations in *Drosophila* suggested that active chromatin exhibits a different IDE than regions characterized by repressive epigenetic marks (Sexton et al., 2012). These IDEs are contrasting our results, as the IDE of heterochromatic PRs showed a higher IDE (−0.7) than euchromatic CAs (−0.85). However, the IDE of repressive epigenetic regions described in *Drosophila* cannot easily be compared to the IDE of constitutive heterochromatin of PRs described in our study. Sexton et al. (2012) pooled various chromatin states, namely, *Polycomb*-silenced chromatin, chromatin bound by Heterochromatin Protein 1, centromeric chromatin, and chromatin that was not enriched in any epigenetic mark (null state). In contrast, the heterochromatin of *Arabidopsis* PRs represents a homogeneous epigenetic state, likely explaining the different IDEs in the two studies.

In accordance with the unchanged SD organization of CAs in *crwn* mutants, the IDEs of CAs in *crwn4* and, to a lesser extent, in *crwn1* resembled IDEs of CAs in the WT. In contrast, the IDEs of PRs were indicative for the FGM and therefore significantly differed from the WT. It is unclear, whether this alteration in the organization of PRs is solely inflicted by reduced nuclear volume or by a function of CRWN4 in centromere organization. Since *crwn1* nuclei are at least as small as *crwn4* nuclei, but disrupted heterochromatic PRs have only been reported in *crwn4* (Wang et al., 2013), the different IDEs of the two mutants in Hi-C experiments support such a function.

In summary, *Arabidopsis* chromosomes are globally organized according to the FGM. However, the PRs are likely organized differently than euchromatic CAs, which can be explained by the fundamentally different roles the two chromatin states play in the nucleus.

The *KNOT* Plays a Role as a Transposon Trap Similar to the *flamenco* Locus in *Drosophila*

As an unexpected, conspicuous feature of the interactome, we observed distinct high IFs between ten *KEE*s, resulting in a web of interactions that we termed *KNOT*. Although *KEE* regions varied among each other with respect to their epigenetic constitution, we observed significant enrichment of associated smRNAs in all *KEE* regions. As *KEE*s were found in fundamentally different chromatin states, such as constitutive heterochromatin of PRs and euchromatic CAs, we did not expect *KEE*s to represent an epigenetically uniform group. By solely considering *KEE*s embedded in euchromatin, we detected an enrichment of H3K27me1 and TEs, suggesting that *KEE*s are heterochromatic islands within euchromatin. However, *KEE* regions are not generally silenced, as actively transcribed genes were detected within them.

Ds transposons preferentially insert in the proximity of *KEE*s. Interestingly, preferential insertion appears to be limited to TEs as we did not observe enriched T-DNA transgene integration near *KEE* regions. The mechanism leading to preferential insertion of TEs within *KEE*s is not solely based on sequence identity of the TEs, as transgenes carrying a *Ds* transposon (WISC lines) were not found to be enriched.

Active TEs potentially harm genome integrity, as TE insertions can disrupt genes and important regulatory elements. Therefore, plants developed a sophisticated TE defense system that relies largely on the RNAi machinery, leading to either posttranscriptional gene silencing or RNA-directed DNA methylation (Castel and Martienssen, 2013). The observed enrichment of new *Ds* insertions and smRNAs, which are associated with *KEE* regions, leads us to propose that the *KNOT* may play a role in TE defense. The *KNOT* might act as a TE trap or relay station from which TEs are either excised or redirected to a TE safe house, such as the PRs.

In *Drosophila*, several piRNA clusters, including the *flamenco* locus, are involved in TE silencing (Brennecke et al., 2007; Malone et al., 2009). Interestingly, *Drosophila* piRNA clusters show similar chromatin interactions as *KEE*s, further supporting the involvement of the *KNOT* in TE defense. Furthermore, it was recently shown that the *flamenco* locus in *Drosophila* serves as a TE trap (Zanni et al., 2013). Based on these similarities, we hypothesize that the *KNOT* is a conserved nuclear structure and plays a similar role as piRNA clusters in *Drosophila*, and that there are nuclear structures analogous to the *KNOT* in other eukaryotes.

Figure 6. The *KNOT*: Epigenetic and Genomic Features and TE Insertion Sites

(A) Distributions of epigenetic and genomic features in *KEE*s (blue) and sampled regions (red). Features that significantly differ in several bin sizes are marked bold.

(B) Interaction between *KEE1* and *KEE6* along 1 Mb. H3K9me2 tracks were 2-fold exaggerated for better visibility.

(C) Interactions among piRNA clusters. Dots represent IFs between piRNA clusters (spanning nine genomic bins of 80 kb each). Boxes indicate IFs of 10,000 randomly sampled regions, selected on chromosomes (ChrX) or CAs (2R, 2L, and 3L), which harbor the respective piRNA clusters. Inset: genomic positions of piRNA clusters in *Drosophila*.

(D) Distribution of natural TE insertion sites (dots) and TE insertion frequencies of RIKEN and CSHL lines (lines).

See also Table S4.

EXPERIMENTAL PROCEDURES

Plant Material

Hi-C experiments were performed using 14-day-old *Arabidopsis thaliana* seedlings (Col-0 accession) grown on Murashige and Skoog culture medium. FISH experiments were performed on Col-0 leaf nuclei. Detailed information on plant materials and growth conditions can be found in the Supplemental Experimental Procedures.

FISH

Chromatin regions encompassing *KEEs* or control regions were hybridized with biotin- or digoxigenin-labeled BAC DNA probes (Table S4). Labeled probes were subsequently detected using secondary antibodies conjugated with fluorescent dyes (Texas Red [red] or Alexa 488 [green]). Pairing events (associations of green and red dots) were subsequently scored using fluorescence microscopy. A detailed protocol for FISH experiments can be found in the Supplemental Experimental Procedures.

Hi-C Experiments

Arabidopsis chromatin was crosslinked and digested using a six-cutter restriction enzyme (HindIII). Subsequently, the chromatin was religated in a large volume favoring intramolecular ligation events, yielding circular DNA molecules comprised of interacting restriction fragments. Identification and quantification of interacting partners were obtained by submitting the DNA to Illumina sequencing, providing genome-wide information on chromosome conformation. A more detailed protocol for Hi-C experiments can be found in the Supplemental Experimental Procedures.

Sequencing reads were aligned to the *Arabidopsis* Col-0 reference genome (TAIR 10) without allowing mismatches and multiple alignments. For subsequent analyses, the mapped sequencing reads were pooled into genomic bins (10, 25, 50, 100, or 250 kb). We then generated matrices in which each element describes the interaction frequency of two genomic bins.

Hi-C Data Analysis

For intrachromosomal interactions, Hi-C matrices were distance normalized by dividing the interaction frequency between two genomic bins by the average interaction frequency of all genomic bins that exhibited the same genomic distance. Subsequently, Pearson's correlation coefficients were calculated such that each element in the correlated Hi-C interaction matrix describes the correlation coefficient between two in silico 4C interactomes (i.e., rows and columns of the distance-normalized interaction matrix). To reveal LSDs and CSDs, respectively, PCA was performed on the correlated Hi-C interaction matrices of single chromosome arms. Genomic bins were then split into two groups according to the sign of the resulting Eigenvalue of each genomic bin (negative Eigenvalues, CSD; positive Eigenvalue, LSD).

To analyze the relationship of chromosome conformation and the epigenetic and genomic landscape, the density (e.g., for histone modifications) or the number (e.g., number of genes) of a given feature per genomic bin was calculated. Based on these values, enrichment or depletion of an epigenetic or genomic feature within LSDs was determined by performing Wilcoxon signed rank testing. Additionally, Pearson's correlation coefficients between the density or count values of a given feature and Eigenvalues of genomic bins was calculated.

SDMs were generated by calculating the sign of the difference between each element of two Hi-C interaction matrices. Subsequently, performing WW testing on each column revealed significant clustering of positive and negative signs, respectively, defining genomic bins that undergo significant architectural changes between the two Hi-C interaction data sets.

To analyze the epigenetic landscape of *KEE* regions and the interaction frequencies between piRNA clusters in *Drosophila,* a Monte-Carlo-based statistical approach was used.

A detailed description of all statistical analyses can be found in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

Hi-C interaction data can be accessed under the Gene Expression Omnibus accession number GSE55960.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2014.07.009.

AUTHOR CONTRIBUTIONS

S.G. and U.G. conceived the study. S.G. performed the experiments. S.G. and M.W.S. performed the bioinformatic data analysis. S.G., M.W.S., and U.G. interpreted the data. S.G., M.W.S., and U.G. wrote the manuscript.

ACKNOWLEDGMENTS

We are indebted to Erika Hughes and Eric J. Richards (Boyce Thomson Institute for Plant Research) for providing seeds of *crwn1* and *crwn4* mutants. We thank Keith Harshman from the Lausanne Genomic Technologies Facility (University of Lausanne) for hospitality during library construction, Eric J. Richards for useful comments on the manuscript, and Konstantinos Kritsas (University of Zürich) for advice on FISH. This work was supported by the University of Zürich, an iPhD project grant from SystemsX.ch, and an Advanced Grant of the European Research Council to U.G.

Received: April 4, 2014
Revised: May 15, 2014
Accepted: July 10, 2014
Published: August 14, 2014

REFERENCES

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., et al. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science *301*, 653–657.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. *2*, 28–36.

Bickmore, W.A., and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*, 1270–1284.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell *128*, 1089–1103.

Castel, S.E., and Martienssen, R.A. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nat. Rev. Genet. *14*, 100–112.

Dittmer, T.A., Stacey, N.J., Sugimoto-Shirasu, K., and Richards, E.J. (2007). *LITTLE NUCLEI* genes affecting nuclear morphology in *Arabidopsis thaliana*. Plant Cell *19*, 2793–2803.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Filion, G.J., van Bemmel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J., and van Steensel, B. (2010). Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. Cell *143*, 212–224.

Fransz, P.F., Armstrong, S., de Jong, J.H., Parnell, L.D., van Drunen, C., Dean, C., Zabel, P., Bisseling, T., and Jones, G.H. (2000). Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: structural organization of heterochromatic *knob* and centromere region. Cell *100*, 367–376.

Fransz, P., De Jong, J.H., Lysak, M., Castiglione, M.R., and Schubert, I. (2002). Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. Proc. Natl. Acad. Sci. USA *99*, 14584–14589.

Grob, S., Schmid, M.W., Luedtke, N.W., Wicker, T., and Grossniklaus, U. (2013). Characterization of chromosomal architecture in *Arabidopsis* by chromosome conformation capture. Genome Biol. *14*, R129.

Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Kleinboelting, N., Huep, G., Kloetgen, A., Viehoever, P., and Weisshaar, B. (2012). GABI-Kat SimpleSearch: new features of the *Arabidopsis thaliana* T-DNA mutant database. Nucleic Acids Res. *40*, D1211–D1215.

Kuromori, T., Hirayama, T., Kiyosue, Y., Takabe, H., Mizukado, S., Sakurai, T., Akiyama, K., Kamiya, A., Ito, T., and Shinozaki, K. (2004). A collection of 11 800 single-copy *Ds* transposon insertion lines in *Arabidopsis*. Plant J. *37*, 897–905.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Luo, C., Sidote, D.J., Zhang, Y., Kerstetter, R.A., Michael, T.P., and Lam, E. (2012). Integrative analysis of chromatin states in *Arabidopsis* identified potential regulatory mechanisms for natural antisense transcript production. Plant J. *73*, 77–90.

Malone, C.D., Brennecke, J., Dus, M., Stark, A., McCombie, W.R., Sachidanandam, R., and Hannon, G.J. (2009). Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. Cell *137*, 522–535.

Marí-Ordóñez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V., and Voinnet, O. (2013). Reconstructing *de novo* silencing of an active plant retrotransposon. Nat. Genet. *45*, 1029–1039.

Mirouze, M., Reinders, J., Bucher, E., Nishimura, T., Schneeberger, K., Ossowski, S., Cao, J., Weigel, D., Paszkowski, J., and Mathieu, O. (2009). Selective epigenetic control of retrotransposition in *Arabidopsis*. Nature *461*, 427–430.

Moissiard, G., Cokus, S.J., Cary, J., Feng, S., Billi, A.C., Stroud, H., Husmann, D., Zhan, Y., Lajoie, B.R., McCord, R.P., et al. (2012). MORC family ATPases required for heterochromatin condensation and gene silencing. Science *336*, 1448–1451.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

Pecinka, A., Schubert, V., Meister, A., Kreth, G., Klatte, M., Lysak, M.A., Fuchs, J., and Schubert, I. (2004). Chromosome territory arrangement and homologous pairing in nuclei of *Arabidopsis thaliana* are predominantly random except for NOR-bearing chromosomes. Chromosoma *113*, 258–269.

Plutarch. (1727). Alexander. In *Plutarch's Lives: Translated from the Greek* (London: J. Tonson).

Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., et al. (2011). Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*. EMBO J. *30*, 1928–1938.

Sakamoto, Y., and Takagi, S. (2013). *LITTLE NUCLEI 1* and *4* regulate nuclear morphology in *Arabidopsis thaliana*. Plant Cell Physiol. *54*, 622–633.

Samson, F., Brunaud, V., Balzergue, S., Dubreucq, B., Lepiniec, L., Pelletier, G., Caboche, M., and Lecharny, A. (2002). FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis thaliana* T-DNA transformants. Nucleic Acids Res. *30*, 94–97.

Schubert, V., Berr, A., and Meister, A. (2012). Interphase chromatin organisation in *Arabidopsis* nuclei: constraints versus randomness. Chromosoma *121*, 369–387.

Sessions, A., Burke, E., Presting, G., Aux, G., McElver, J., Patton, D., Dietrich, B., Ho, P., Bacwaden, J., Ko, C., et al. (2002). A high-throughput *Arabidopsis* reverse genetics system. Plant Cell *14*, 2985–2994.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the *Drosophila* genome. Cell *148*, 458–472.

Sundaresan, V., Springer, P., Volpe, T., Haward, S., Jones, J.D., Dean, C., Ma, H., and Martienssen, R. (1995). Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. Genes Dev. *9*, 1797–1810.

Wang, H., Dittmer, T.A., and Richards, E.J. (2013). *Arabidopsis* CROWDED NUCLEI (CRWN) proteins are required for nuclear size control and heterochromatin organization. BMC Plant Biol. *13*, 200.

Woody, S.T., Austin-Phillips, S., Amasino, R.M., and Krysan, P.J. (2007). The WiscDsLox T-DNA collection: an *Arabidopsis* community resource generated by using an improved high-throughput T-DNA sequencing pipeline. J. Plant Res. *120*, 157–165.

Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., Luyten, I., Quesneville, H., Vaury, C., and Jensen, S. (2013). Distribution, evolution, and diversity of retrotransposons at the *flamenco* locus reflect the regulatory properties of piRNA clusters. Proc. Natl. Acad. Sci. USA *110*, 19842–19847.

Zhang, Y., McCord, R.P., Ho, Y.-J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell *148*, 908–921.

8.3 Apomictic and Sexual Germline Development Differ with Respect to Cell Cycle, Transcriptional, Hormonal and Epigenetic Regulation

The following manuscript is published in "PLOS Genetics" (open access)[1]. I helped analyzing the data, provided tools and code for data analysis, and performed some of the transcriptome analysis (mainly the part on the annotation of the *de novo* assembled transcriptome and raw data pre-processing). I also isolated egg cells and synergids of *Arabidopsis thaliana* for transcriptional profiling. I further helped to write a part of the manuscript (methods section).

[1]Schmidt, A, Schmid, MW, Klostermeier, UC, Qi, W, Guthörl, D, Sailer, C, Waller, M, Rosenstiel, P, and Grossniklaus, U (2014) Apomictic and Sexual Germline Development Differ with Respect to Cell Cycle, Transcriptional, Hormonal and Epigenetic Regulation. PLOS Genetics 10: e1004476.

OPEN ACCESS Freely available online

PLOS | GENETICS

Apomictic and Sexual Germline Development Differ with Respect to Cell Cycle, Transcriptional, Hormonal and Epigenetic Regulation

Anja Schmidt[1]*, Marc W. Schmid[1], Ulrich C. Klostermeier[2], Weihong Qi[3], Daniela Guthörl[1], Christian Sailer[1], Manuel Waller[1], Philip Rosenstiel[2], Ueli Grossniklaus[1]*

1 Institute of Plant Biology & Zürich-Basel Plant Science Center, University of Zürich, Zürich, Switzerland, **2** Institute of Clinical Molecular Biology, Christian-Albrechts University, Kiel, Germany, **3** Functional Genomics Center Zürich, UZH/ETH Zürich, Zürich, Switzerland

Abstract

Seeds of flowering plants can be formed sexually or asexually through apomixis. Apomixis occurs in about 400 species and is of great interest for agriculture as it produces clonal offspring. It differs from sexual reproduction in three major aspects: (1) While the sexual megaspore mother cell (MMC) undergoes meiosis, the apomictic initial cell (AIC) omits or aborts meiosis (apomeiosis); (2) the unreduced egg cell of apomicts forms an embryo without fertilization (parthenogenesis); and (3) the formation of functional endosperm requires specific developmental adaptations. Currently, our knowledge about the gene regulatory programs underlying apomixis is scarce. We used the apomict *Boechera gunnisoniana*, a close relative of *Arabidopsis thaliana*, to investigate the transcriptional basis underlying apomeiosis and parthenogenesis. Here, we present the first comprehensive reference transcriptome for reproductive development in an apomict. To compare sexual and apomictic development at the cellular level, we used laser-assisted microdissection combined with microarray and RNA-Seq analyses. Conservation of enriched gene ontologies between the AIC and the MMC likely reflects functions of importance to germline initiation, illustrating the close developmental relationship of sexuality and apomixis. However, several regulatory pathways differ between sexual and apomictic germlines, including cell cycle control, hormonal pathways, epigenetic and transcriptional regulation. Enrichment of specific signal transduction pathways are a feature of the apomictic germline, as is spermidine metabolism, which is associated with somatic embryogenesis in various plants. Our study provides a comprehensive reference dataset for apomictic development and yields important new insights into the transcriptional basis underlying apomixis in relation to sexual reproduction.

Citation: Schmidt A, Schmid MW, Klostermeier UC, Qi W, Guthörl D, et al. (2014) Apomictic and Sexual Germline Development Differ with Respect to Cell Cycle, Transcriptional, Hormonal and Epigenetic Regulation. PLoS Genet 10(7): e1004476. doi:10.1371/journal.pgen.1004476

Editor: Imran Siddiqi, CCMB, United States of America

Received November 15, 2013; **Accepted** March 18, 2014; **Published** July 10, 2014

Funding: This project was supported by the University of Zürich, the Marie Curie project IDEAGENA (to AS), and grants from the "Staatssekretariat für Bildung und Forschung" in the framework of COST action FA0903 (to AS and UG) and the Swiss National Science Foundation (to UG). PR was supported by the DFG Cluster of excellence Inflammation at Interfaces (CL Nucleotide lab). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: aschmidt@botinst.uzh.ch (AS); grossnik@botinst.uzh.ch (UG)

Introduction

In flowering plants, both sexual and asexual reproduction through seeds (apomixis) is common. Apomixis occurs in more than 400 plant species belonging to over 40 families, but it is poorly represented in crop species. Apomixis leads to clonal offspring by conservation of the maternal genotype through the absence of meiosis and fertilization [1–4]. Engineering of apomixis in crop species is perceived as one of the greatest challenges faced by modern agriculture [5]. However, achieving this goal proved to be difficult, particularly as the knowledge about the genetic basis and regulatory programs underlying apomictic reproduction is very limited.

Sexual reproduction and apomixis only differ in a number of key developmental steps [6,7]. During sexual reproduction, the female and male reproductive lineages are initiated by spore formation from a spore mother cell during megasporogenesis and microsporogenesis, respectively. The megaspore mother cell (MMC) is the first cell of the female germline. It is specified by selection of one subepidermal, somatic (sporophytic) cell within an ovule, the precursor of the seed. The MMC undergoes meiosis and gives rise to a tetrad of reduced megaspores. Typically, only one of those - the functional megaspore (FMS) - survives to form the female gametophyte (embryo sac). The FMS divides mitotically and subsequently cellularizes to form the mature female gametophyte harbouring the gametes (egg cell and central cell) and several accessory cells, including the synergids that play an important role in fertilization [8]. Double fertilization of the egg and central cell with one sperm each initiates the development of embryo and endosperm, respectively. In contrast, in gametophytic apomixis an unreduced sporophytic cell of the ovule in proximity to the MMC (apospory), or the MMC itself becoming an apomictic initial cell (AIC) that omits or aborts meiosis (diplospory), gives rise to an unreduced embryo sac (apomeiosis) [9]. The egg cell subsequently develops into an embryo without fertilization (parthenogenesis). Endosperm development can either be autonomous or require fertilization (pseudogamy). It is likely that signals from sporophytic ovule tissues are important for the

Author Summary

In flowering plants, asexual reproduction through seeds (apomixis) likely evolved from sexual ancestors several times independently. Only three key developmental steps differ between sexual reproduction and apomixis. In contrast to sexual reproduction, in apomicts the first cell of the female reproductive lineage omits or aborts meiosis (apomeiosis) to initiate gamete formation. Subsequently, the egg cell develops into an embryo without fertilization (parthenogenesis), and endosperm formation can either be autonomous or depend on fertilization. Consequently, the offspring of apomicts is genetically identical to the mother plant. The production of clonal seeds bears great promise for agricultural applications. However, the targeted manipulation of reproductive pathways for seed production has proven difficult as knowledge about the underlying gene regulatory processes is limited. We performed cell type-specific transcriptome analyses to study apomictic germline development in *Boechera gunnisoniana*, an apomictic species closely related to *Arabidopsis thaliana*. To facilitate these analyses, we first characterized a floral reference transcriptome. In comparison, we identified several regulatory pathways, including core cell cycle regulation, protein degradation, transcription factor activity, and hormonal pathways to be differentially regulated between sexual and apomictic plants. Apart from new insights into the underlying transcriptional networks, our dataset provides a valuable starting point for functional investigations.

development of the sexual and apomictic germline [6,9]. During meiosis the MMC is shielded by incorporation of callose into its cell wall [10], which may temporarily reduce or block such signaling. However, to our knowledge such signaling events have so far not been investigated in detail.

While recent studies uncovered the transcriptional basis of key steps of female germline development in the sexual model species *Arabidopsis thaliana* [11–13], relatively little is known about the genetic and transcriptional basis governing apomictic reproduction. Gametophytic apomixis is genetically controlled by usually two or more loci - or potentially clusters of linked loci - in different aposporous and diplosporous species [14–24]. In the *Boechera* genus, there is evidence for a complex genetic control of apomixis [25]. At the transcriptional level it has been hypothesized that apomixis is derived from a deregulation of the sexual pathway [6,7,26]. Indeed, evidence for differential regulation of many genes between apomictic and sexual accessions comes from comparative gene expression analyses. These studies mostly use ovule or flower tissues from a variety of species, including *Boechera* spp. [27,28], *Brachiaria* spp. [29,30], *Hieracium perforatum* [31], *Pennisetum* spp. [32,33], *Paspalum* spp. [34–36], apomeiotic mutants of *Medicago falcata* [37], *Panicum maximum* [38], and *Poa pratensis* [39,40]. In addition, recent findings indicate spatial and temporal shifts in the expression of genes important for reproductive development between sexual and apomictic plants [27–29,41].

To coordinate such complex transcriptional deregulation, the involvement of epigenetic regulatory pathways has been proposed [3,6,7]. Epigenetic pathways play important roles in regulating developmental and cell-fate decisions through the modification of gene activity by histone modifications, DNA methylation or gene silencing by small RNAs. Interestingly, features of apospory or diplospory have recently been observed in *Arabidopsis* and maize carrying mutant alleles of genes involved in DNA methylation and small RNA pathways [42–44]. In *Arabidopsis* plants carrying mutations in *ARGONAUTE9* (*AGO9*), or genes encoding additional members of a small RNA pathway (*RNA-DEPENDENT RNA POLYMERASE 6* (*RDR6*), *SUPPRESSOR OF GENE SILENCING 3* (*SGS3*)), additional MMC-like cells in the ovule gave rise to developing female gametophytes in a process resembling apospory [42]. Maize plants with mutations in homologues of the *Arabidopsis* DNA methyltransferases *DOMAINS REARRANGED METHYLASE1* (*DRM1*)/*DRM2* and *CHROMOMETHYLTRANSFERASE3* (*CMT3*) show also features of apospory [43]. However, in maize plants carrying mutations in *AGO104*, a homologue of *Arabidopsis AGO9*, formation of unreduced viable gametes occurs by a diplospory-like mechanism [44].

In addition, features of apospory have been observed in *Arabidopsis* plants carrying mutations in the RNA helicase gene *MNEME* (*MEM*), which restricts germline fate to one cell per ovule [12]. As in *ago9* mutants, the additional MMC-like cells initiate development of unreduced female gametophytes [12]. Apomeiosis has also been achieved by mutating important meiotic genes in *Arabidopsis*, such as *DYAD/SWITCH1* (*SWI1*), a regulator of meiotic chromosome organisation, or a combination of three mutations in the MiMe triple mutant (*sporulation11-1* (*spo11-1*); *omission of second division1* (*osd1*); *recombination8* (*rec8*)) [45,46]. However, to date the potential role of these genes in naturally occurring apomixis has not been elucidated.

To study the transcriptional basis of key steps of apomictic reproduction we used the triploid, diplosporous species *Boechera gunnisoniana* as an apomictic model. The genus *Boechera* is closely related to the sexual model species *Arabidopsis thaliana*, facilitating comparative studies. We demonstrate the obligate apomictic behaviour of *B. gunnisoniana* by analysing the ploidy of embryo and endosperm in single seeds by means of a flow cytometric seed screen [47]. As no annotated, genome-wide sequence information is available for this species, we used RNA-Seq (Illumina HiSeq2000) to generate a reference transcriptome based on ovule tissues isolated by microdissection at the developmental stages of interest. We annotated the reference transcriptome, including the identification of homologous genes in *Arabidopsis*. Using a combination of laser-assisted microdissection (LAM), Affymetrix GeneChip profiling (ATH1), and RNA-Seq (SOLiD), we studied the transcriptome of isolated AICs, as well as egg, central and synergid cells from *B. gunnisoniana*. Statistical data analysis revealed the significant enrichment of polyamine and spermidine metabolism in the AIC as compared to the cells of the mature female gametophyte in *Boechera*. In addition, we compared the gene expression profiles of the AIC and the MMC, egg cells and central cells between apomictic *Boechera* and sexual *Arabidopsis*. This uncovered differential expression of genes in important regulatory pathways, including protein degradation, hormonal pathways, cell cycle control, signal transduction, transcriptional regulation, and epigenetic pathways.

Results

Boechera gunnisoniana seeds are derived from unreduced female gametes

B. gunnisoniana has previously been described as diplosporous apomict [48,49]. While the embryo develops parthenogenetically, the endosperm requires fertilization (pseudogamy) [48,49]. Based on flow cytometric studies of single seeds, a high variability of the reproductive mode - ranging from obligate sexual to obligate apomictic - has been reported among 71 *Boechera* accessions analysed [50]. We applied this technique to test the frequency of apomictic reproduction in *B. gunnisoniana*. From 84 individual

seeds tested, ~98% showed a 3C:9C (embryo:endosperm) ploidy ratio in the seed, as expected for a triploid, pseudogamous apomict (Figure 1A). In two seeds (~2%) a 6C embryo resulted from fertilization of an unreduced egg cell (Figure 1B). In conclusion, *B. gunnisoniana* reproduces obligatory by pseudogamous apomixis. In all seeds analysed an unreduced egg cell gave rise to the embryo, and embryos developed parthenogenetically at very high frequency.

Nevertheless, the possibility of developmental variations during germline formation cannot be excluded based on a flow cytometric analysis alone. We used ovule and seed clearings for cytological analyses to address the question whether there is potential variation of reproductive development. In young ovules typically a single enlarged subepidermal cell specified to an AIC (Figure S1A, B), while in 3.6% of all ovules (N = 551) an additional enlarged, subepidermal cell was observed (Figure S1A). As previously reported, the AICs give rise to the formation of dyads [48,49,51]. Dyad formation was seen at a frequency of 85% (N = 224; Figure S1E, Q). In an additional 10% of all ovules, either dyads accompanied by large parietal cells and or triads were formed (Figure S1F, Q). These two possibilities could not clearly be discriminated based on morphology. Unexpected numbers of nuclei during AIC division or the formation tetrads were observed in ~2% of all cases (Figure S1G, Q). In the remaining 3% of ovules the AICs apparently failed to divide (Figure S1C, Q), likely leading to developmental arrest (Figure S1D). Formation of a mature gametophyte was observed in 92% of all ovules (N = 353) in agreement with previously published results [49], the majority showing a delay or defect in the fusion of the polar nuclei (Figure S1I, J, R). In 7.4% of the ovules development was arrested early (at AIC or FMS stage), was delayed, or resulted in an unexpected number of nuclei (Figure S1R). At a very low frequency (0.6%) more than one gametophyte developed in a single ovule (Figure S1K, R). In agreement with previous reports, in the absence of pseudogamous fertilization no evidence for the initiation of embryo development was observed [48,51]. After fertilization, 62% of the seeds developed normally (N = 477; Figure S1L, M). In the remainder, ovules harbouring mature gametophytes or enlarging seeds due to seed coat growth without embryo or endosperm development were observed, or only embryo or endosperm development initiated (Figure S1N–P), suggesting a problem in fertilization. In summary, in *B. gunnisoniana* the large majority of mature gametophytes are formed by diplospory and 98% of the seeds are derived parthenogenetically under our growing conditions. Thus *B. gunnisoniana* is well suited as a model species for molecular studies of apomixis.

Sequencing, assembly, and annotation of a *Boechera gunnisoniana* reference transcriptome based on ovule tissues

The close relation of the apomict *B. gunnisoniana* with the sexual model species *A. thaliana* provides an excellent basis for comparative analyses. However, while genome sequencing projects for *Boechera* species are currently ongoing (http://www.jgi.doe.gov/), this initiative does not include *B. gunnisoniana*, which is fast cycling and obligatory diplosporous. Thus, as a tool for transcriptomic studies, we generated a reference transcriptome for this species. We isolated ovules at the two developmental stages of interest, megasporogenesis (i.e. ovule stages from the initiation of integument development until the integuments start to overgrow the nucellus; Figure S1A, B) and mature gametophyte stage (Figure S1I–K). The highly enriched ovule samples included some pistil tissue, particularly for the early developmental stage. We prepared two libraries that were sequenced with the Illumina HiSeq2000. Both libraries were assembled together using trinity [52]. Following removal of reads with low average quality scores (Q<30) or adaptor sequences, and trimming of low quality (Q< 20) ends, around 697 million reads were assembled into 112'232 sequences corresponding to 30'298 distinct genes with 50% having a sequence length of ≥2'153 bp. The reference transcriptome was annotated using Blast2GO [53] and BLAT [54] (Table S1). Using Blast2GO, 51% of all hits matched best to *A. thaliana* and an

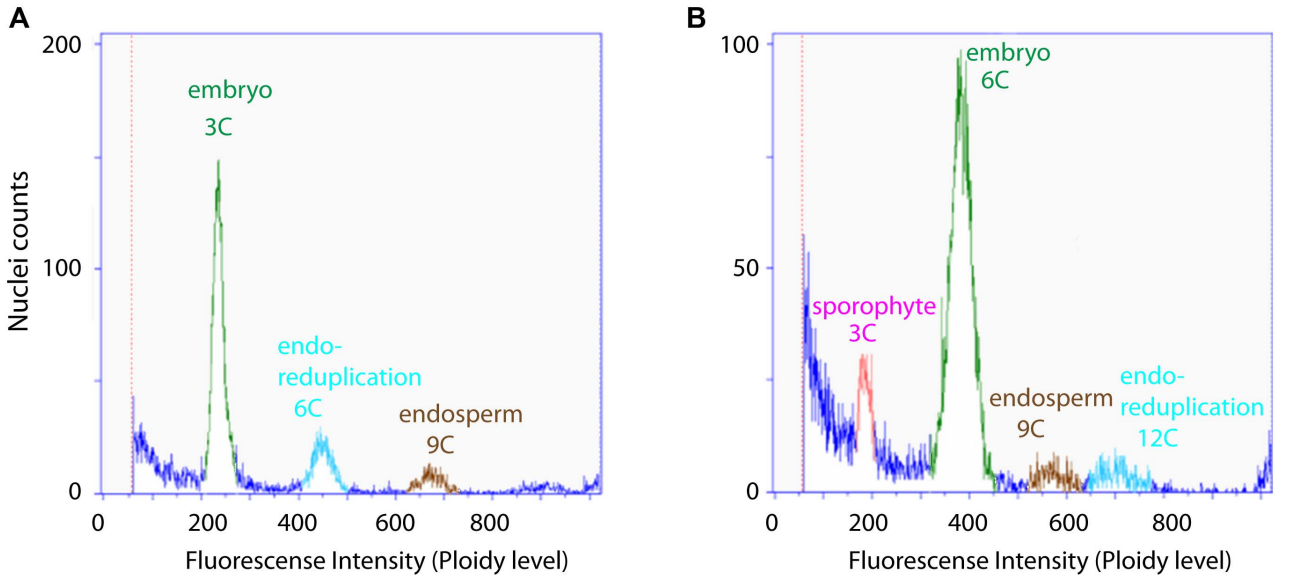


Figure 1. Flow cytometric seed screen on single seeds of *Boechera gunnisoniana* to analyse ploidy. Fluorescence intensity was measured on individual green seeds of *Boechera* to determine the ploidy level of embryo and endosperm. 98% of the seeds measured (N = 84) showed a 3C:9C ratio of embryo:endosperm, indicating diplosporous apomixis (A). The remaining 2% showed a 6C:9C embryo:endosperm ratio indicative of a BIII hybrid where the embryo is derived from fertilization of an unreduced egg cell (B).
doi:10.1371/journal.pgen.1004476.g001

additional 25% to *A. lyrata* sequences. Gene ontology (GO) terms could be successfully assigned to 62% of all hits. In addition, we aligned the sequences to cDNA (TAIR10) using BLAT and identified 19'617 close *A. thaliana* homologues of *B. gunnisoniana* genes (hereafter denoted as *Arabidopsis* homologues, Table S2). In summary, the length of assembled sequences and annotation results indicate a good quality of our apomictic reference transcriptome.

Transcriptional profiling of cells involved in key steps of gametophytic apomixis

For the sexual model plant *Arabidopsis*, transcriptomes of the cell types of the mature gametophyte (egg, central, and synergid cells) and the MMC have been described [11–13]. From these studies, important new insights into the transcriptional basis of sexual germline development could be gained. We applied LAM to isolate the AIC and the surrounding sporophytic nucellus tissue, as well as the egg, central, and synergid cells from *B. gunnisoniana* (Figure 2A, B; Figure S2A). For the AIC, small contamination with surrounding nucellus tissue cannot be completely avoided (Figure 2A, B). Due to the dense structure of the mature embryo sac, samples for egg, central, and synergid cells are highly enriched in these cell types, but contain some contamination from neighbouring gametophytic cells (Figure S2A). For transcriptional profiling, 300–650 cell- or tissue-specific sections were pooled per sample. Transcriptional profiling was done using two alternative strategies: heterologous hybridization of amplified and labelled *Boechera* RNA to the Affymetrix ATH1 GeneChip designed for *Arabidopsis* and SOLiD V4 sequencing (Table 1, Figure 2C, D). For GeneChip analysis, the extracted RNA was subjected to linear amplification, labelled and hybridized to the microarray as described [12]. Cross-species hybridization of microarrays with RNA from a species other than the original target species is largely influenced by the degree of sequence similarities between the probes on the array and the mRNA sequence of the species under investigation [55]. To account for this effect we used an adapted *Bg*PANP algorithm for the generation of presence/absence p values, similar to the *At*PANP previously shown to outperform the default algorithm [11]. These algorithms use probes that do not match to the reference genome or transcriptome of the target species as "negative probes" to estimate the true background of each array. For our *Bg*PANP algorithm probes not aligning to the reference transcriptome (allowing for three mismatches) were defined as negative. In this way, several thousand genes were detected significantly above background (hereafter referred to as present/"P") in each cell type-specific sample (Table 1, Figure 2C, Figure S2B, C, Table S3). For RNA-Seq, the isolated RNA was subjected to linear amplification following an established protocol [13,56]. Each library was sequenced on one eights of a slide, resulting in 53'701'313 (AIC, apo_initial3), 50'453'327 (egg cell, egg_cell2), 49'331'759 (central cell, central_cell2), and 46'240'916 (synergid cells, synergid_cell2) reads. Reads were processed and aligned to the assembled reference transcriptome as described [13]. Under the applied criteria, between 30% and 37% of the reads had at least one valid alignment, corresponding to 16'371'464 (apo_initial3), 18'783'550 (egg_cell2), 17'348'718 (central_cell2), and 15'353'384 (synergid_cell2) weighted alignments. Gene expression values were calculated as the sum of expression of individual variants (Table S4). We identified 16'385, 17'828, 19'091, and 10'409 *B. gunnisoniana* genes to be expressed (i.e. to have at least 5 mapped reads) in the AIC, egg, central, and synergid cells, respectively (Table 1). This corresponds to 13'047, 13'811, 14'893, and 9'390 expressed (≥5 read counts) *Arabidopsis* homologues in the AIC, egg, central, and synergid cells, respectively (Table 1, Table S2, Table S4). Between ~2'000 and 6'000 genes were consistently identified in at least two independent cell type-specific samples (Table 1), in agreement with previous observations on the comparability of microarray and RNA-Seq data and the higher sensitivity and genome-wide coverage reached by RNA-Seq [13].

Independent data confirmation shows apomictic initial cell-enriched expression

Apomixis and sexual reproduction are interrelated developmental processes. Therefore, it is likely that the cell type-specific transcriptome profiles are largely overlapping between the sexual and apomictic mode of reproduction. Nevertheless, differences in

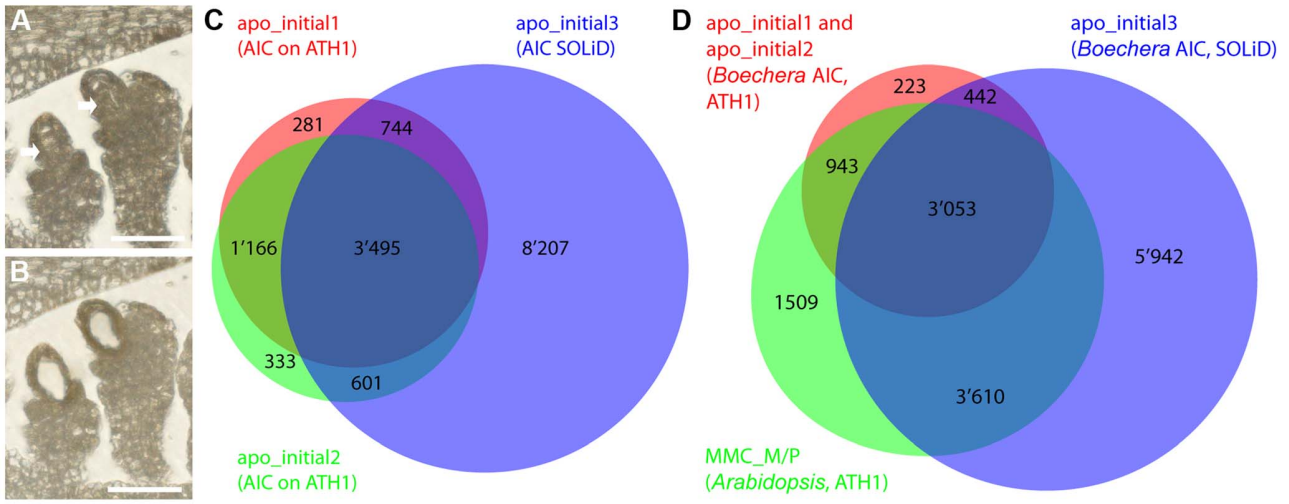


Figure 2. Laser-assisted microdissection (LAM) and transcriptome analysis to study the *Boechera* apomictic initial cell (AIC). (A, B) LAM of the AIC from a 6 µm dry section (scale bar = 40 µm). (A) An ovule harbouring the AIC before LAM. Arrows point to the AICs. (B) The ovule after the AIC has been dissected and collected. (C, D) Venn diagrams showing the overlaps of prediction of expression (P calls; apo_initial1, apo_initial2, MMC_M/P) as determined with the *Bg*PANP algorithm or the *At*PANP algorithm described previously [12], and the genes with ≥5 read counts on *Boechera* homologues (apo_initial3) to the *Arabidopsis* genes as determined by mapping to the *Boechera* reference transcriptome.
doi:10.1371/journal.pgen.1004476.g002

expression of a subset of genes are expected due to the differences in reproductive mode and species. To compare the cell type-specific transcriptome profiles between *Boechera* and *Arabidopsis*, we used genes designated as P in two (for AIC) or one (for egg and central cell) microarray sample(s), or were identified as an expressed *Arabidopsis* homologue using RNA-Seq (Table 1). For *Arabidopsis* we used the 9'115 genes with evidence of expression in the MMC [12], 12'769 genes expressed in the egg cell (as described in [11,12] and SOLiD reads aligned to the reference genome of *Arabidopsis thaliana* (TAIR10)), and 14'661 genes expressed in the central cell ([11,12] and both samples from [13]). Comparing the genes with evidence of expression from *Arabidopsis* and *Boechera* for MMC/AIC, egg and central cells, we found overlapping expression of 7'606, 9'883, and 10'772 genes, respectively (Figure 2C, D; Figure S2B, C).

In addition, we selected several genes for independent data confirmation by *in situ* hybridization. Based on our analyses, these genes were expressed in the *Boechera* AIC but not in the *Arabidopsis* MMC (Table S5). Probes for *in situ* hybridization on *B. gunnisoniana* ovule sections were designed based on the *Arabidopsis* Col-0 cDNA for three transcription factors (Figure 3, (A–D) AT1G06170, basic helix-loop-helix (bHLH) DNA-binding superfamily protein; (E–G) AT1G28050, *B-BOX DOMAIN PROTEIN 13*; (H–J) AT1G76580, Squamosa promoter-binding protein-like (SBP domain) transcription factor family protein), an oligopeptide transporter (AT1G59740, Figure 3 K,L), and a HIGH MOBILITY GROUP A protein (HMGA, AT1G14900, Figure 3 M–O). The probes were designed to have significant sequence homologies only to the respective *Boechera* homologue (Figure S3, Supporting Information S1). For all selected genes we could confirm enriched expression in the AIC. Taken together, our analyses confirm the accuracy of the *B. gunnisoniana* transcriptome dataset.

Gene expression and gene ontology enrichment analysis uncovers upregulation of spermidine metabolism in the apomictic initial cell

Between sexual and apomictic reproduction, there are important differences in cell specification and cell fate decisions. Heterochronic shifts in expression patterns have been reported previously using isolated *Boechera* ovules from sexual and apomictic accessions [27,28]. However, gene expression has not yet been profiled in a germline-specific manner without the confounding effects of the surrounding sporophytic tissue in *Boechera*. Based on genes significantly enriched in the MMC as compared to the cell types of the mature gametophyte, we previously identified translational control pathways and the activity of RNA-helicases as crucial for the acquisition of germline fate and MMC specification in *Arabidopsis* [12]. To see if similar or different functions are prominent in the *Boechera* AIC as compared to the mature gametophyte, we used read counts obtained by mapping to the *Boechera* reference transcriptome. To identify genes significantly enriched we used NOIseq-sim, a non-parametric approach for differential expression analysis based on simulated replicate samples [57]. We identified 1'487 genes to be significantly enriched in the AIC as compared to the cell types of the mature gametophyte (Figure 4A). In addition, 3'509, 1'466, and 1'806 genes were significantly enriched in the egg, central, and synergid cells, respectively, as compared to the other three cell types of the germline under investigation.

In a gene ontology (GO) analysis, we identified functions important for pollen germination and sperm cell and pollen maturation as significantly enriched in the AIC ($p<0.01$, Table 2). In addition, different metabolic and transport processes were upregulated, in addition to spermidine metabolism and polyamine biosynthesis ($p<0.01$, Table 2). Functions related to plant cell wall modification and epigenetic regulatory pathways (histone H3K4 demethylation and maintenance of DNA methylation) were also amongst the enriched functions ($p<0.01$, Table 2). Furthermore, cytokinin catabolism was among the near-significantly enriched processes ($p = 0.012$, Table 2).

In the egg cell of *Boechera*, cytokinin metabolism is a dominant molecular function as discovered by analysis of GO enrichment based on the 3'509 significantly upregulated genes, in addition to transcription factor activity (Table S6A). The central cell transcriptome is dominated by different epigenetic regulatory pathways, cell cycle regulation, and regulation of cell fate decisions ($p<0.01$, Table S6B).

Table 1. Transcriptome analysis of 11 samples from apomictic *Boechera* isolated by LAM.

Sample	Genes present on microarray with *Bg*PANP (p value≤0.02)	Number of *B. gunnisoniana* genes with ≥5 read counts	Number of *Arabidopsis* homologues with ≥5 read counts	Expressed in at least 2 samples (column 1 and column 3)
apo_initial1	5'595			6'006
apo_initial2	5'686			
apo_initial3		16'385	13'047	
sporo_nucellus1	5'706			4'345
sporo_nucellus2	5'236			
egg_cell1	5'835			4'490
egg_cell2		17'828	13'811	
central_cell1	2'973			2'192
central_cell2		19'091	14'893	
synergid_cell1	4'149			2'472
synergid_cell2		10'409	9'390	

Summary of gene expression from *Boechera* germline samples. Samples apo_initial1 and 2, sporo_nucellus1 and 2, egg_cell1, central_cell1, and synergid_cell1 were hybridized on ATH1 microarrays, apo_initial3, egg_cell2, central_cell2, and synergid_cell2 were analysed using RNA-Seq (SOLiD V4) by mapping of reads to the *B. gunnisoniana* reference transcriptome.

doi:10.1371/journal.pgen.1004476.t001

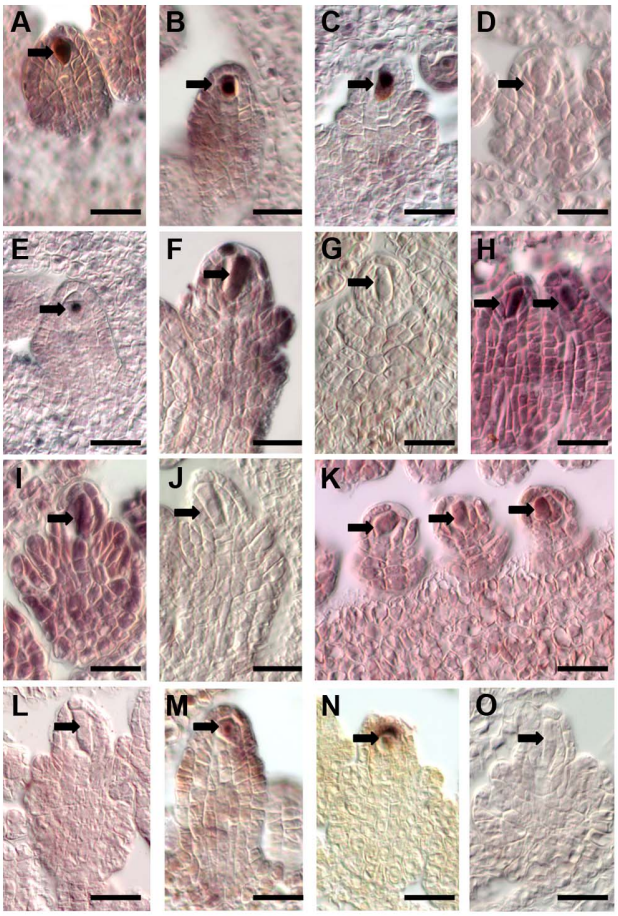


Figure 3. Independent data validation for selected genes by *in situ* hybridization on *B. gunnisoniana* ovules. Data validation for selected genes found expressed in the *Boechera* AIC but not the *Arabidopsis* MMC. Scale bars are 20 µm, arrows point to the AICs. *In situ* hybridizations on *B. gunnisoniana* ovule sections were performed with antisense probes (A–C, E, F, H, I, K, M, N) or sense probes as controls (D, G, J, L, O) for the transcription factors AT1G06170, a basic helix-loop-helix (bHLH) DNA-binding superfamily protein (A–D), AT1G28050, a *B-BOX DOMAIN PROTEIN 13* (E–G), and AT1G76580 a Squamosa promoter-binding protein-like (SBP domain) transcription factor family protein (H–J), an oligopeptide transporter, AT1G59740 (K,L), and AT1G14900, encoding the HIGH MOBILITY GROUP A protein.
doi:10.1371/journal.pgen.1004476.g003

At higher stringency, using EdgeR with an estimated biological variation coefficient of 0.8, we identified 142 genes to be significantly enriched in all pairwise comparisons of the AIC with the transcriptomes of cells of the mature gametophyte (adjusted p value (FDR)<0.05, Benjamini-Hochberg adjustment; Figure 4B) [58]. Based in these genes, GO enrichment analysis confirmed spermidine metabolism, cytokinin catabolism, and functions related to pollen development and germination as significantly enriched in the AIC ($p<0.01$; Table S7). Notably, also the term "sexual reproduction" was an enriched function based on upregulated genes. In addition, 3'792 genes were differentially expressed in any pairwise comparison between the cell types of the mature gametophyte (FDR≤0.05 for comparisons between synergid cells and egg- or central cell, or an unadjusted p value≤0.001 for comparisons between egg cell and central cell (Figure 4B)).

In summary this indicates interesting differences in the functions underlying the specification of the germline lineage and the female gametes in the apomict *B. gunnisoniana* as compared to the sexual pathway in *Arabidopsis*. Consistently, spermidine metabolism was identified as enriched in the AIC. Our analysis also indicated a distinct regulation of cytokinin metabolism and degradation in the apomictic germline lineage.

Evidence for different regulation of important regulatory pathways in apomictic and sexual germline cells

To analyse differences in gene activity between the sexual and apomictic germline in more detail, we identified *Arabidopsis* genes and their homologues only expressed in a certain cell type in *Arabidopsis* or *Boechera*. *Boechera* genes were designated as expressed when having at least 5 read counts by mapping against the reference transcriptome, or a P call on one or both microarrays. For a conservative estimate of genes only expressed in *Arabidopsis*, we also aligned the SOLiD reads to the reference genome of *A. thaliana* (TAIR10) and only considered genes with at least 5 read counts. We included the latter method as annotation of the closest *Arabidopsis* homologue is not always unambiguous. Sometimes sequence variants for one *Boechera* gene have their highest sequence similarity to different *Arabidopsis* genes (see below), complicating a direct comparison. Of the 9'115 MMC-expressed genes, no evidence of expression has been found for 852 genes in the AIC. GO analysis on this set of genes identified a significant enrichment of different molecular functions, including metabolism, regulation of physiological responses, auxin turnover, translation initiation, and functions related to cell wall structure and cell cycle control ($p<0.01$, Table 3A). Also the "core cell cycle genes" were found to be significantly enriched (Fisher's exact test, $p = 0.006$), in agreement with the meiotic fate of the MMC. In addition, 14 protein family (PFAM) domains were identified as enriched (Fisher's exact test, p value<0.01, Table S8) including F-box domain and F-box related domains, as well as the cyclin C- and N-terminal domains. This suggests that protein ubiquitinylation and degradation, as well as cell cycle control, may be differentially regulated between MMCs and AICs. Using a similar approach, out of 12'679 genes expressed in the *Arabidopsis* egg cell (Figure S2B) we identified 1'731 for which no homologues were expressed in the *Boechera* egg cell. GO analysis in this set of genes identified biological processes related to RNA modification and splicing, transport and metabolism, and methylation-dependent chromatin silencing as significantly enriched, and also functions related to double fertilization and endosperm formation ($p<0.01$, Table 4A). In addition, two transcription factor families, the "*At*RKD Transcription Factor Family" and the "MYB Transcription Factor Family" were identified as significantly enriched gene families (Fisher's exact test, $p = 0.0087$ and $p = 0.0038$, respectively). For the *Arabidopsis* central cell, out of 14'661 expressed genes (Figure S2C) no evidence for expression of homologues in the *Boechera* central cell was found for 2'146 genes. As in the *Arabidopsis* egg cell, biological processes related to RNA modification and splicing (GO:0000154, rRNA modification, $p = 5.1e\text{-}17$; GO:0045292, nuclear mRNA *cis*-splicing, via spliceosome, $p = 5.4e\text{-}5$) and "endosperm development" ($p = 0.0078$) were significantly enriched. In addition, out of the 12 PFAM domains identified as enriched, three were related to F-box domains (Fisher's exact test, $p<0.01$, Table S9).

For the identification of genes only expressed in the apomictic *Boechera* germline and not in *Arabidopsis*, we used the *Arabidopsis* homologues identified and mapping to the *Boechera* reference transcriptome, combined with the microarray data. We identified 5'273 and 4'902 genes expressed in the apomictic egg and central cell, respectively, that were absent in the corresponding *Arabidopsis* cell type. We used more restrictive criteria to identify the

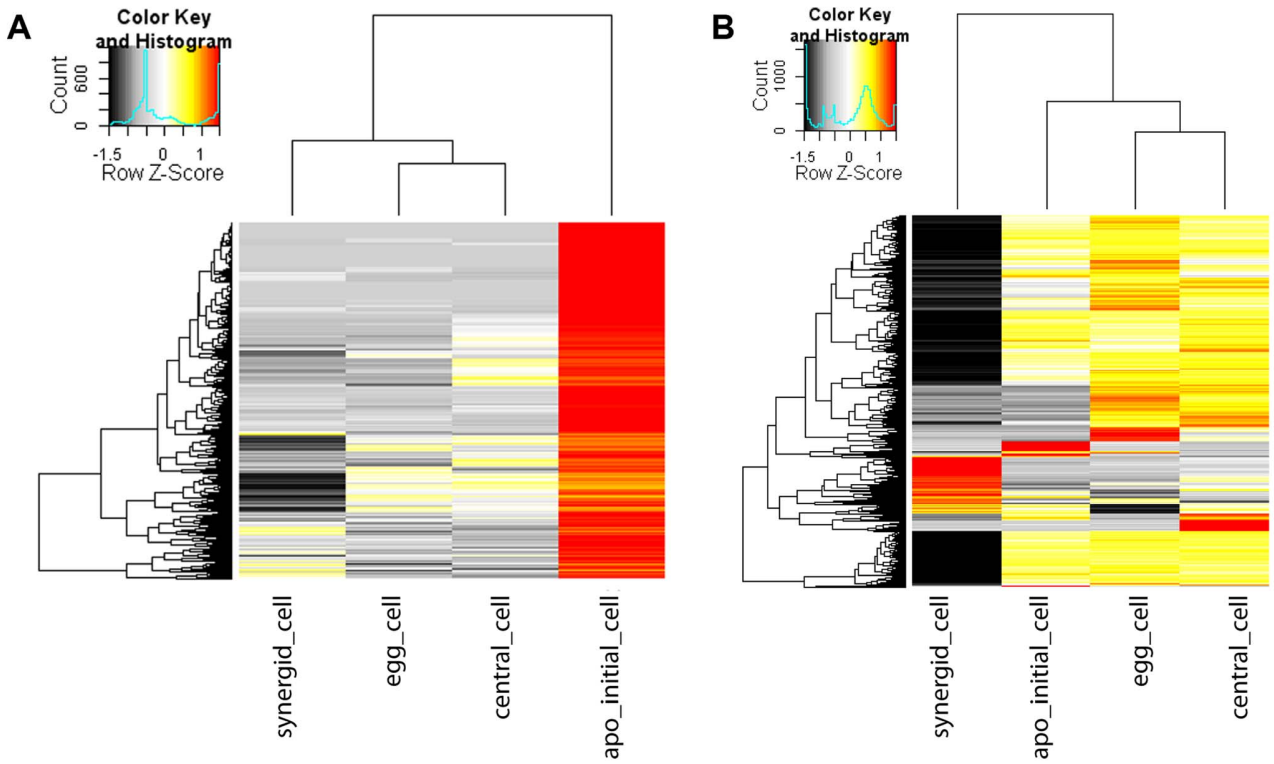


Figure 4. Heatmap of log2 transformed normalized read counts. Heatmap of 1′487 genes enriched in the *Boechera* AIC as compared to all cell types of the mature female gametophyte as identified using NOISeq-sim (A). Heatmap of 3′792 genes enriched in the AIC as compared to all cell types of the mature gametophyte or differentially expressed between any cell type of the mature gametophyte identified using EdgeR (B). The hierarchical clustering of samples and genes was based on euclidean distance and hierarchical agglomerative clustering. Colours are scaled per row and red denotes high expression and black low expression.
doi:10.1371/journal.pgen.1004476.g004

901 genes expressed in the AIC but not in the MMC: we considered only *Arabidopsis* homologues with ≥5 reads in the SOLiD dataset and detected as P in at least one microarray dataset of the AIC. Interestingly, for all three cell types of the apomictic *Boechera* germline, GO and/or PFAM analyses revealed a significant enrichment of signal transduction processes and protein kinases (Table 3B, Table 4B, Table 5, Table S10). For instance, we identified the significant enrichment of "MAP kinase kinase activity" in the AIC ($p<0.01$, Table 3B). In addition, transport and metabolic processes were enriched, and spermidine metabolism was confirmed as an important feature ($p<0.01$, Table 3B). Analysis of gene families revealed the Squamosa promoter Binding Proteins as enriched (Fisher's exact test, p value<0.01, SBP transcription factor family). Analysis of gene families and PFAM domains also identified a significant enrichment of the *AMINO ACID/AUXIN PERMEASE* (*AAAP*) family, the *ARF* transcription factor family, and the protein domain of the AUX/IAA family during apomictic germline specification (Fisher's exact test, $p<0.01$, Table 6). Also the family of B3 transcription factors (B3_TFs), including the *ARF* transcription factor family, was identified as significantly enriched (Fisher's exact test, $p<0.01$, Table 6). In the parthenogenetic *Boechera* egg cell, GO analysis suggests the importance of signal transduction pathways, cell cycle regulation, and transcription factor activity ($p<0.01$). In general, in the female gametes analysis of gene families identified the enriched expression of several transcription factor families, particularly the basic Helix-Loop-Helix transcription factors both in the egg and central cell (Table S11).

In summary, our analysis reveals interesting differences in the regulatory programs underlying the acquisition of germline fate and between the female gametes. While the subset of genes only expressed in the sexual germline is significantly enriched in protein degradation pathways, the apomictic *Boechera* germline is marked by the activity of signal transduction processes. In addition, indications for a role of auxin signalling and metabolism were observed in both germlines. Among the genes identified as active in the apomictic germline lineage only, we found the enrichment of different transcription factors families, particularly basic helix-loop-helix transcription factors in the female *Boechera* gametes. The comparison between the sexual and apomictic germlines further revealed differential regulation of genes involved in cell cycle control and posttranscriptional regulatory processes, including mRNA splicing, and epigenetic regulatory pathways related to methylation-dependent chromatin modifications.

Expression analysis of selected candidate genes and pathways related to apomixis

For a number of genes, enriched expression in the *Arabidopsis* MMC or the aposporic initial cell of *Hieracium praealtum* have previously been described [12,31]. In addition, for sexual or apomictic germline development, evidence for the importance of different genes including core cell cycle genes, meiotic genes, and

Table 2. Gene ontology analysis.

GO.ID	Term	Annotated	Significant	Expected	p value
GO:0010584	pollen exine formation	92	44	4.73	<1e-30
GO:0009827	plant-type cell wall modification	336	46	17.26	1.20E-09
GO:0008216	spermidine metabolic process	25	10	1.28	2.00E-07
GO:0009860	pollen tube growth	393	43	20.19	2.50E-06
GO:0006527	arginine catabolic process	8	5	0.41	1.70E-05
GO:0006817	phosphate ion transport	32	9	1.64	2.30E-05
GO:0008643	carbohydrate transport	97	18	4.98	2.30E-05
GO:0046467	membrane lipid biosynthetic process	210	25	10.79	8.00E-05
GO:0006596	polyamine biosynthetic process	24	7	1.23	0.00015
GO:0016036	cellular response to phosphate starvation	194	22	9.97	0.00042
GO:0055085	transmembrane transport	1020	80	52.4	0.00048
GO:0034720	histone H3-K4 demethylation	5	3	0.26	0.00125
GO:0009396	folic acid-containing compound biosynthetic process	18	5	0.92	0.00173
GO:0030162	regulation of proteolysis	6	3	0.31	0.0024
GO:0048235	pollen sperm cell differentiation	47	8	2.41	0.00248
GO:0046938	phytochelatin biosynthetic process	7	3	0.36	0.00405
GO:0006665	sphingolipid metabolic process	88	12	4.52	0.00582
GO:0030036	actin cytoskeleton organization	254	14	13.05	0.00587
GO:0010216	maintenance of DNA methylation	15	4	0.77	0.00599
GO:0009395	phospholipid catabolic process	15	4	0.77	0.00599
GO:0010951	negative regulation of endopeptidase activity	8	3	0.41	0.00623
GO:0010199	organ boundary specification between lateral organs and meristems	8	3	0.41	0.00623
GO:0015800	acidic amino acid transport	7	3	0.36	0.00763
GO:0010084	specification of organ axis polarity	3	2	0.15	0.00764
GO:0090408	phloem nitrate loading	3	2	0.15	0.00764
GO:0042398	cellular modified amino acid biosynthetic process	84	10	4.32	0.01076
GO:0010205	photoinhibition	18	4	0.92	0.01188
GO:0009823	cytokinin catabolic process	10	3	0.51	0.01236

Biological Processes identified to be up-regulated based on 1′487 genes identified to be up-regulated in the AIC as compared to the cell types composing the mature gametophyte (egg cell, central cell, synergid cells).
doi:10.1371/journal.pgen.1004476.t002

genes involved in epigenetic regulatory pathways has previously been reported based on mutant analyses or expression patterns [42,43,59]. Thus, we compared the expression of selected genes of interest upon sexual and diplosporic germline initiation.

From a list of 89 core cell cycle genes as defined before [60,61], 75 are represented on the ATH1 array and for 66 *Arabidopsis* genes, homologues were identified in the *Boechera* reference transcriptome. From these, 41 genes are expressed in the *Arabidopsis* MMC and 49 homologues are present in the *Boechera* AIC. 16 and 24 cell cycle regulators have only been detected in the MMC or the AIC, respectively (Table S12). In particular, the genes only detected in the apomict upon germline specification include genes involved in different cell cycle transitions, e.g. G1/S phase, including a number of genes involved in the cyclin D/retinoblastoma/E2F pathway (Table S12). The observed differences in cell cycle regulation are in agreement with the different mechanisms of cell division in the meiotic MMC *versus* the diplosporous AIC.

Interestingly, for 14 selected meiotic genes and genes expressed in the sexual MMC no evidence for expression was found in the *H. praealtum* aposporous initial cell [31]. However, although the aposporous and the diplosporous initial cell both give rise to unreduced embryo sacs, cell lineage and developmental fate are markedly different. So far it is unknown whether common regulators underlie apomeiosis in these distinct types of apomixis. Interestingly, for all 14 genes except for *SWITCH/DYAD* and *SPO11-2* evidence for expression was found in the *Boechera* AIC; although at very low levels for most genes (Figure 5). The *Arabidopsis* male meiocytes cluster separately from the expression data of different *Arabidopsis* cell- and tissue-types publicly available [13,62–66]. Furthermore, the RNA helicase *MEM* previously identified as predominantly expressed in the *Arabidop-*

Table 3. Gene ontology analysis on MMC and AIC.

A) *Arabidopsis thaliana* MMC

GO.ID	Term	Annotated	Significant	Expected	p value
GO:0016538	cyclin-dependent protein kinase regulator activity	29	6	1.14	0.00081
GO:0005199	structural constituent of cell wall	30	6	1.18	0.00098
GO:0016844	strictosidine synthase activity	14	4	0.55	0.00176
GO:0004129	cytochrome-c oxidase activity	15	4	0.59	0.00232
GO:0010178	IAA-amino acid conjugate hydrolase activity	3	2	0.12	0.00455
GO:0003743	translation initiation factor activity	82	9	3.24	0.00496
GO:0005034	osmosensor activity	4	2	0.16	0.00885

B) *Boechera gunnisoniana* AIC

GO.ID	Term	Annotated	Significant	Expected	p value
GO:0005275	amine transmembrane transporter activity	69	14	2.94	1.40E-05
GO:0022843	voltage-gated cation channel activity	28	6	1.19	0.00098
GO:0050662	coenzyme binding	321	27	13.66	0.00121
GO:0072547	tricoumaroylspermidine meta-hydroxylase activity	2	2	0.09	0.00181
GO:0072548	dicoumaroyl monocaffeoyl spermidine meta-hydroxylase activity	2	2	0.09	0.00181
GO:0072549	monocoumaroyl dicaffeoyl spermidine meta-hydroxylase activity	2	2	0.09	0.00181
GO:0016620	oxidoreductase activity, acting on the aldehyde or oxo group of donors	42	7	1.79	0.00181
GO:0005315	inorganic phosphate transmembrane transport activity	14	4	0.6	0.00232
GO:0016844	strictosidine synthase activity	14	4	0.6	0.00232
GO:0004091	carboxylesterase activity	109	12	4.64	0.00232
GO:0016160	amylase activity	17	4	0.72	0.00498
GO:0000257	nitrilase activity	3	2	0.13	0.00528
GO:0016298	lipase activity	107	11	4.55	0.00586
GO:0004708	MAP kinase kinase activity	11	3	0.47	0.00981

Significant upregulation of molecular functions based on 852 genes expressed in the *Arabidopsis* MMC but not in the *Boechera* AIC (A). Significant upregulation of molecular functions based on 901 genes expressed in the apomictic initial cell but not in the MMC (B). A p value$<$0.01 was considered significant.
doi:10.1371/journal.pgen.1004476.t003

sis MMC is only expressed at low levels in the AIC but higher in the apomictic egg cell (Figure 5). Interestingly, this indicates differences in the expression of genes previously identified to have important functions for MMC specification and meiosis in *Arabidopsis*. In agreement with the differences in developmental fate, the data also suggest differences in cell specification of aposporous and diplosporous initial cells. Nevertheless, the majority of 35 genes described as enriched in the *H. praealtum* aposporous initial cell or the early apomictic embryo sac as compared to sporophytic ovule tissues [31] was also expressed in the *Boechera* AIC, except for *HISTONE ACETYLTRANSFERASE OF THE CBP FAMILY1*, *LIKE HETEROCHROMATIN PROTEIN1*, *BEL1-LIKE HOMEODOMAIN1*, *CONSTITUTIVE DISEASE RESISTANCE1*, genes involved in lipid localization (*AT1G03103*, *AT5G38170*, *AT3G18280*, *AT1G43666*), and a pathogenesis-related lipid-transfer protein gene (*AT2G18370*).

Increasing evidence suggests the involvement of epigenetic regulatory pathways in the discrimination between sexual reproduction and apomixis. Therefore, we were interested in a closer investigation of the expression of 69 genes involved in DNA methylation and small RNA pathways (as used in [12]). 58 of these genes have annotated homologues in *Boechera* (Table S1). 40 genes are consistently present both in the AIC and in the MMC, supporting the important role of epigenetic regulatory pathways for the initiation of germline development [12]. Heatmap clustering suggests the closest relation between the AIC dataset and the datasets of the *Boechera* female gametes (Figure S4). Together, these datasets cluster with the *Arabidopsis* egg and synergid cells, but distantly from male meiocytes or the central cell of the sexual germline lineage (Figure S4). Nevertheless, a number of genes were only detected in the MMC or the AIC, respectively (Supporting Information S2). Genes only detected in the AIC included *ENHANCED SILENCING PHENOTYPE3* (*ESP3*). Also *AGO9* and *RDR6*, mutations in which cause an apospory-like behaviour in *Arabidopsis* [42], were both detected at low levels in the *Boechera* AIC (Figure S4, Figure S5). In summary, for a subset of genes involved in DNA

Table 4. Gene ontology analysis on sexual and parthenogenetic egg cells of *Arabidopsis* and *Boechera*, respectively.

(A) ***Arabidopsis thaliana*** **egg cell**

GO.ID	Term	Annotated	Significant	Expected	p value
GO:0000154	rRNA modification	74	43	4.02	<1e-30
GO:0045292	nuclear mRNA cis splicing, via spliceosome	8	4	0.43	0.00051
GO:0045490	pectin catabolic process	2	2	0.11	0.00295
GO:0015986	ATP synthesis coupled proton transport	38	7	2.07	0.00395
GO:0043086	negative regulation of catalytic activity	82	11	4.46	0.00474
GO:0019432	triglyceride biosynthetic process	7	3	0.38	0.00475
GO:0080155	regulation of double fertilization forming a zygote and endosperm	3	2	0.16	0.00854
GO:2000014	regulation of endosperm development	3	2	0.16	0.00854
GO:0090309	positive regulation of methylation-dependent chromatin silencing	3	2	0.16	0.00854

(B) ***Boechera gunnisoniana*** **egg cell**

GO.ID	Term	Annotated	Significant	Expected	p value
GO:0006468	protein phosphorylation	1135	308	199.51	2.70E-15
GO:0006355	regulation of transcription, DNA-dependent	1868	421	328.36	2.20E-08
GO:0007169	transmembrane receptor protein tyrosine kinase signalling pathway	130	49	22.85	4.00E-08
GO:0007623	circadian rhythm	100	33	17.58	0.00014
GO:0006952	defense response	884	201	155.39	0.00016
GO:0009739	response to gibberellin stimulus	123	42	21.62	0.00043
GO:0009641	shade avoidance	13	8	2.29	0.0005
GO:0019747	regulation of isoprenoid metabolic process	10	7	1.76	0.00059
GO:0007165	signal transduction	1319	314	231.85	0.00067
GO:0006974	response to DNA damage stimulus	228	61	40.08	0.00068
GO:0006261	DNA-dependent DNA replication	65	23	11.43	0.00092
GO:0010321	regulation of vegetative phase change	4	4	0.7	0.00095
GO:0010215	cellulose microfibril organization	4	4	0.7	0.00095
GO:0000079	regulation of cyclin-dependent protein serine/threonine kinase activity	46	17	8.09	0.00139
GO:0009314	response to radiation	570	131	100.19	0.00209
GO:0010158	abaxial cell fate specification	7	5	1.23	0.00257
GO:0009737	response to abscisic acid stimulus	412	95	72.42	0.0026
GO:0016998	cell wall macromolecule catabolic process	30	12	5.27	0.00321
GO:2000038	regulation of stomatal complex development	10	6	1.76	0.00323
GO:0010065	primary meristem tissue development	10	6	1.76	0.00323
GO:0006571	tyrosine biosynthetic process	5	4	0.88	0.0041
GO:0009939	positive regulation of gibberellic acid mediated signalling pathway	5	4	0.88	0.0041
GO:0046482	para-aminobenzoic acid metabolic process	5	4	0.88	0.0041
GO:0042372	phylloquinone biosynthetic process	8	5	1.41	0.00585
GO:0010162	seed dormancy process	21	9	3.69	0.0061
GO:0010200	response to chitin	127	34	22.32	0.0063
GO:0009826	unidimensional cell growth	211	52	37.09	0.00764

Table 4. Cont.

(B) *Boechera gunnisoniana* egg cell					
GO.ID	**Term**	**Annotated**	**Significant**	**Expected**	**p value**
GO:0006928	cellular component movement	93	26	16.35	0.00871
GO:0016556	mRNA modification	15	7	2.64	0.00889
GO:0048868	pollen tube development	116	31	20.39	0.00903
GO:0009694	jasmonic acid metabolic process	29	11	5.1	0.00957
GO:0006499	N-terminal protein myristoylation	431	95	75.76	0.00974
GO:0006857	oligopeptide transport	76	22	13.36	0.00986

Significant enrichment of biological processes based on 1'731 genes with evidence of expression only in the sexual *Arabidopsis* egg cell (A) and 5'273 *Boechera* homologues with evidence of expression only in the parthenogenetic egg cell (B). A p value$<$0.01 was considered significant.
doi:10.1371/journal.pgen.1004476.t004

methylation and small RNA pathways, we observed distinct expression patterns during germline specification in sexual *Arabidopsis* MMCs *versus* apomictic *Boechera* AICs, which may be of importance to distinguish fate decisions between these alternative reproductive modes.

Influence of sequence similarities between *Arabidopsis* and *Boechera* homologues on the distribution of count data

Particularly within a gene family, the assignment of the closest *Boechera* homologues to *Arabidopsis* genes is not always unambiguous. For selected gene families of interest we aimed to test the influence of sequence divergence and annotation criteria on the expression estimates for *B. gunnisoniana* homologues of *A. thaliana* genes. Identification of the closest homologues in the *Boechera* reference transcriptome was based on the highest bit score sum with BLAT, using only the best mappings per *Arabidopsis* gene. For this analysis we selected the *AtRKD* gene family (Figure 6, 7). In addition, similar analysis of the *ARIADNE* (*ARI*) gene family, and the *AGO* gene family are shown in ''Supporting Information S2'' (Figure S5, Figure S6, Figure S7, Supporting Information S2).

The *RKD* gene family has been identified in our analysis to be enriched among the genes expressed only in the *Arabidopsis* but not the *Boechera* egg cell. Instead of the five members of the *Arabidopsis RKD* family, two gene models of homologues with one variant each have been identified in the *Boechera* reference transcriptome (Figure 6). This suggests either that the gene family is smaller in *Boechera* as compared to *Arabidopsis*, or that additional members of this family are not expressed in *Boechera* ovules at the developmental stages used to generate the reference transcriptome. Analysis of sequence similarities indicates the closest similarity between comp76373_c0_seq1 and *AtRKD2*. In agreement, counts for reads mapped to comp76373_c0_seq1 are assigned to *AtRKD2*. However, while clustering of comp83606_c0_seq1 indicates higher sequence divergence from all *AtRKDs*, the reads are assigned to *AtRKD5*. The expression and role of members of the *RKD* family in *Arabidopsis*, where they play a role in egg cell specification, has been described previously [67]. As the two *Boechera* gene models homologous to the *AtRKD* genes are expressed in the egg apparatus (egg and synergid cells; Figure 6, 7), the *Arabidopsis* family as a whole is predominantly expressed in the *Arabidopsis* egg apparatus (Figure 7), in agreement with our gene set enrichment analysis.

Discussion

Boechera gunnisoniana as a model species to study apomixis

To investigate apomictic reproduction, the female germline is in particular of interest, as in apomicts clonal offspring genetically identical to the mother plant is generated. In *B. gunnisoniana*, based on a flow cytometric seed screen using single seeds, we observed exclusively apomeiotic behaviour and only a very low percentage of fertilized, unreduced egg cells. In agreement, the formation of dyads and mature *Polygonum* type embryo sacs were observed at high frequencies. At low frequencies, developmental variations during germline development were observed, including the formation of more than one female gametophyte per ovule. This could either be due to a failure of degradation of the second megaspore resulting from diplospory, or indicate the rare occurrence of apospory. Interestingly, parthenogenesis remains repressed in the absence of pseudogamous fertilization. In maturing siliques, likely due to a lack of successful fertilization, not all female gametes give rise to an embryo or endosperm. As a consequence of deviations from apomictic germline development and fertilization, reproductive development seems to arrests, so that the vast majority of mature seeds are derived apomictically. This obligate apomictic behaviour, together with its fast cycling (about 4 months from seed to seed) and the close relation to the sexual model species *A. thaliana*, make *B. gunnisoniana* an ideal system to study apomixis. We generated the first comprehensive, annotated reference transcriptome for reproductive development in *B. gunnisoniana*, including the identification of *Arabidopsis* homologues, as an essential tool for further studies.

Spermidine and polyamine metabolism are enriched in the apomictic initial cell

Previously, similarities of germline development were reported even across kingdoms, between the plant and animal germline. These are likely of general importance for the acquisition of germline fate [12]. Nevertheless, cell type specification and developmental fate is markedly different during germline specification in sexual, aposporous, and diplosporous species. Consistently, a number of differences in gene expression profiles have been observed between the apomictic and the sexual germline. In the *B. gunnisoniana* AIC, a number of functions related to pollen development and germination were enriched, consistent with gene activities observed during germline development in apomeiotic, non-parthenogenetic hybrids of *Pennisetum glaucum* [33].

Table 5. Analysis of PFAM domains significantly enriched in the *Boechera* germline.

AIC				
ID	**Significant**	**Expected**	**p value**	**description**
PF01535	217	124.07	2.30E-10	Pentatricopeptide repeat
PF00069	268	200.47	0.00012525	Protein kinase domain
PF00225	38	18.94	0.00123647	Kinesin motor domain
PF00612	32	16.10	0.00295969	IQ calmodulin-binding motif
PF07714	130	96.92	0.00671958	Protein typrosine kinase
egg cell				
ID	**Significant**	**Expected**	**p value**	**description**
PF00010	42	24.88	0.00476216	Helix-loop-helix DNA-binding domain
PF00069	183	131.90	0.00016034	Protein kinase domain
PF00560	131	73.15	6.30E-08	Leucine rich repeat
PF00730	9	2.25	0.00288164	HhH-GPD superfamily base excision DNA repair protein
PF00786	10	2.99	0.004385	P21-Rho binding domain
PF00931	68	31.43	5.85E-07	NB-ARC domain
PF01535	197	82.32	3.67E-21	Pentacortico repeat
PF01582	42	24.88	0.00476216	Toll Interleukin receptor
PF07646	12	3.74	0.00238746	Kelch motif
PF07714	100	63.80	0.00017157	Protein tyrosine kinase
PF07725	34	15.34	0.00027842	Leucine rich repeat, LRR_3
PF08263	75	39.10	5.75E-06	Leucine rich repeat N-terminal domain
central cell				
ID	**Significant**	**Expected**	**p value**	**description**
PF00010	48	26.1526398	0.00058139	Helix-loop-helix DNA-binding domain
PF00069	222	138.628655	7.03E-09	Protein kinase domain
PF00076	23	48.1759154	0.00025011	RNA recognition motiv
PF00225	26	11.9948198	0.00198709	Kinesin motor domain
PF00560	137	76.8848283	4.57E-08	Leucine rich repeat
PF00612	24	11.0116378	0.00227686	IQ calmodolin binding domain
PF00931	64	33.0349134	2.48E-05	NB-ARC domain
PF01486	19	8.06209197	0.00435678	K-box domain
PF01535	191	86.5200114	1.47E-17	Pentacorticopeptide repeat
PF01582	44	26.1526398	0.00422265	Toll-interleukin receptor
PF02183	12	4.12936418	0.00722254	Homeobox associated leucine zipper
PF02362	33	18.2871842	0.00530936	B3 domain
PF02458	24	12.3880925	0.00850195	Transferase family
PF03000	16	6.48900085	0.00559973	NPH3-family
PF03514	17	6.48900085	0.00323279	GRAS domain family
PF04570	11	3.14618223	0.00233681	DUF581
PF04770	13	3.34281862	0.00047255	ZF-HD protein dimerisation region
PF04784	10	3.14618223	0.0059545	DUF547
PF07714	115	67.0530088	2.29E-06	Tyrosine kinase
PF07725	36	16.1241839	0.00015028	Leucine rich repeat, LRR3
PF08263	84	41.0970054	2.01E-07	Leucine rich repeat N-terminal domain
PF11721	16	6.29236446	0.005044	Di-glucose binding with endoplasmatic reticulum

PFAM domains significantly enriched in 901, 5'273 and 4'902 genes expressed only in the AIC, egg cell and central cell of *Boechera* but without evidence of expression in the *Arabidopsis* MMC, egg cell, and central cell, respectively (p value$<$0.01).

doi:10.1371/journal.pgen.1004476.t005

Table 6. Gene family enrichment.

Gene family	Significant	Expected	p value
AAAP family	8	1.55	0.0004025
Acyl Lipid Metabolism Family	35	20.65	0.0050592
ARF Transcription Factor Family	5	0.71	0.0015891
B3_TFs	8	2.77	0.0099967
Glycoside Hydrolase Gene Families	24	13.30	0.0084674
Monolignol Biosynthesis	8	2.31	0.0037947
Organic Solute Cotransporters	28	10.20	7.53E-06
SBP Transcription Factor Family	4	0.59	0.0051551
Superfamily of zinc-coordinating DNA-binding proteins	4	0.59	0.0051551

Enrichment of gene families in 901 genes with evidence of expression in the AIC but not in the sexual MMC as analysed by Fisher's exact test. A p value<0.01 was considered significant.
doi:10.1371/journal.pgen.1004476.t006

Polyamine biosynthesis and spermidine metabolism were also identified as features of the *Boechera* AIC. Interestingly, spermidine synthesis is essential for embryo development in *Arabidopsis* [68]. In addition, a possible role of polyamines in promoting somatic plant embryogenesis was reported [69–71]. This indicates the importance of spermidine for plant reproduction and provides an interesting link between polyamines and somatic embryogenesis, a form of asexual reproduction different

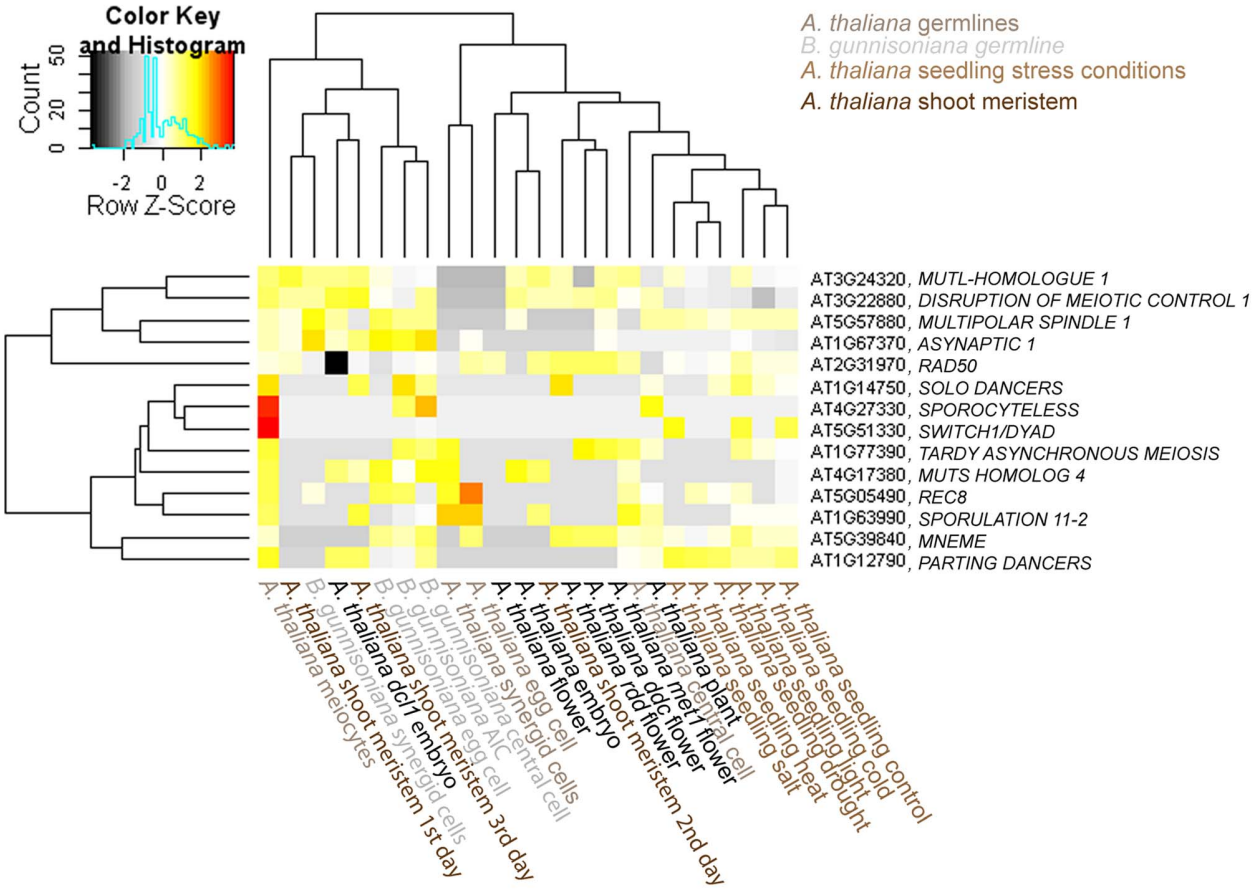


Figure 5. Heatmap of log2 transformed normalized read counts for 14 selected meiotic or MMC-expressed genes. Hierachical clustering of read counts from different *Arabidopsis* and *Boechera* cell- and tissue types [13,62–66]. The hierarchical clustering of samples and genes was based on euclidean distance and hierarchical agglomerative clustering. Colours are scaled per row. Red denotes high expression and black low expression.
doi:10.1371/journal.pgen.1004476.g005

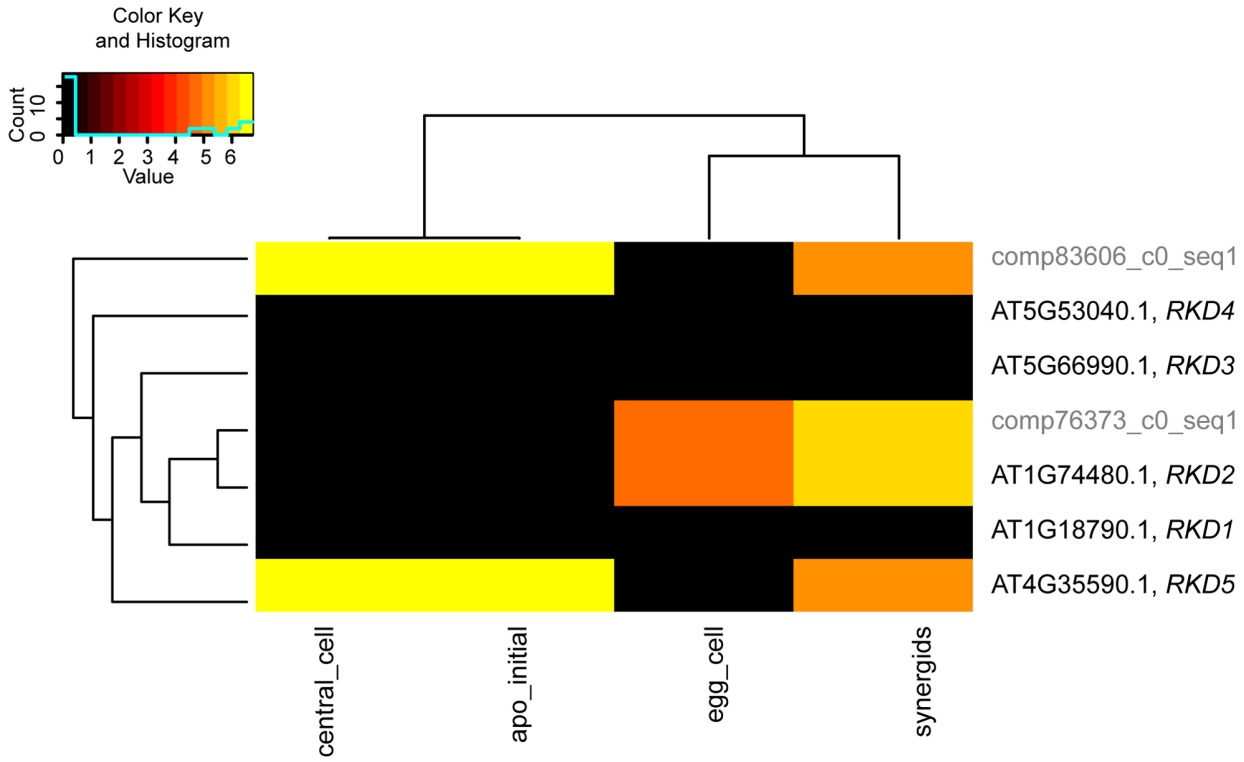


Figure 6. Analysis of sequence divergence of members of the *AtRKD* gene family. Analysis of sequence divergence of members of the *AtRKD* gene family and close *Boechera* homologues as analysed with ClustalX based on protein sequences and read counts assigned. *Boechera* gene model variants are indicated with compxxx_cY_seqZ.
doi:10.1371/journal.pgen.1004476.g006

from gametophytic apomixis. Interestingly, spermidine is involved in the protection of DNA from oxidative stress by quenching free radicals mostly arising from reactive oxygen species (ROS) [72]. In line with the high activity of spermidine metabolism in the apomeiotic AIC, it has been hypothesized that repair of DNA damage after oxidative stress has been a major driving force for the evolution of meiosis [73]. Apart from being cytotoxic, the role of ROS in signalling and for plant reproductive development has recently been demonstrated [74]. Notably, a spermine/spermidine synthase has previously been identified to be present in the apospory-specific region of *P. squamulatum* and hypothesized to be expressed [75], supporting a potential role of these substances for the specification of the apomictic germline. However, further studies will be required to conclude which, if any, role polyamine and spermidine metabolism plays during germline development or the determination of the asexual reproductive fate.

Differentially regulated genes and pathways during sexual and apomictic reproduction include hormonal and protein degradation pathways and transcription factor activity

In addition to polyamine and spermidine metabolism, the activities of important hormonal pathways were also observed in the AIC. Upregulation of cytokinin degradation was detected upon apomictic germline specification as compared to the mature gametophyte, while the egg cell is marked by gene activities leading to cytokinin modifications. In addition, genes involved in auxin signalling were enriched in the set of genes expressed in the AIC but not in the sexual MMC, in line with the identification of genes involved in auxin signal transduction in the *H. praealtum* apospory initial cell [31]. In the *Boechera* AIC, we detected an enriched activity of the AUX/IAA and the ARF transcription factor gene families. These play crucial roles in auxin-regulated gene expression, for example to control cell type-specific auxin responses during *Arabidopsis* embryo development [76,77]. Evidence for differential expression of ARF genes has previously been reported during early stages of reproductive development in a comparative cDNA-AFLP analysis targeting sexual and apomictic *Paspalum simplex* flowers [35].

In contrast, genes active only during sexual reproduction and MMC specification are marked by an enrichment of F-box proteins. F-box proteins play important roles in ubiquitin-dependent protein degradation involved in signal transduction pathways, cell cycle control, and a variety of other processes [78,79]. The expression of miRNAs targeting genes encoding F-box proteins and ARF transcription factors in *Boechera* floral tissues supports the importance of these pathways in plant reproductive development [80]. This is in line with the identification of a truncated *ARI* allele with homology to *Arabidopsis ARI7* as a candidate apospory locus in *Hypericum perforatum* [81]. *ARI7* encodes a ring finger protein predicted to be involved in ubiquitin-dependent protein degradation [82]. Interestingly, we found evidence for higher activity of *ARI* family members in the sexual MMC compared to the AIC.

In addition to miRNAs targeting F-box proteins and ARF transcription factors, miRNAs involved in regulation of SPL and MYB transcription factors have been identified in *Boechera* spp. [80]. Together with the enrichment observed for SPL transcription factors in the *B. gunnisoniana* AIC and of MYB transcription factors in the sexual *Arabidopsis* MMC and egg cell, respectively, this suggests that these transcription factors play important roles in

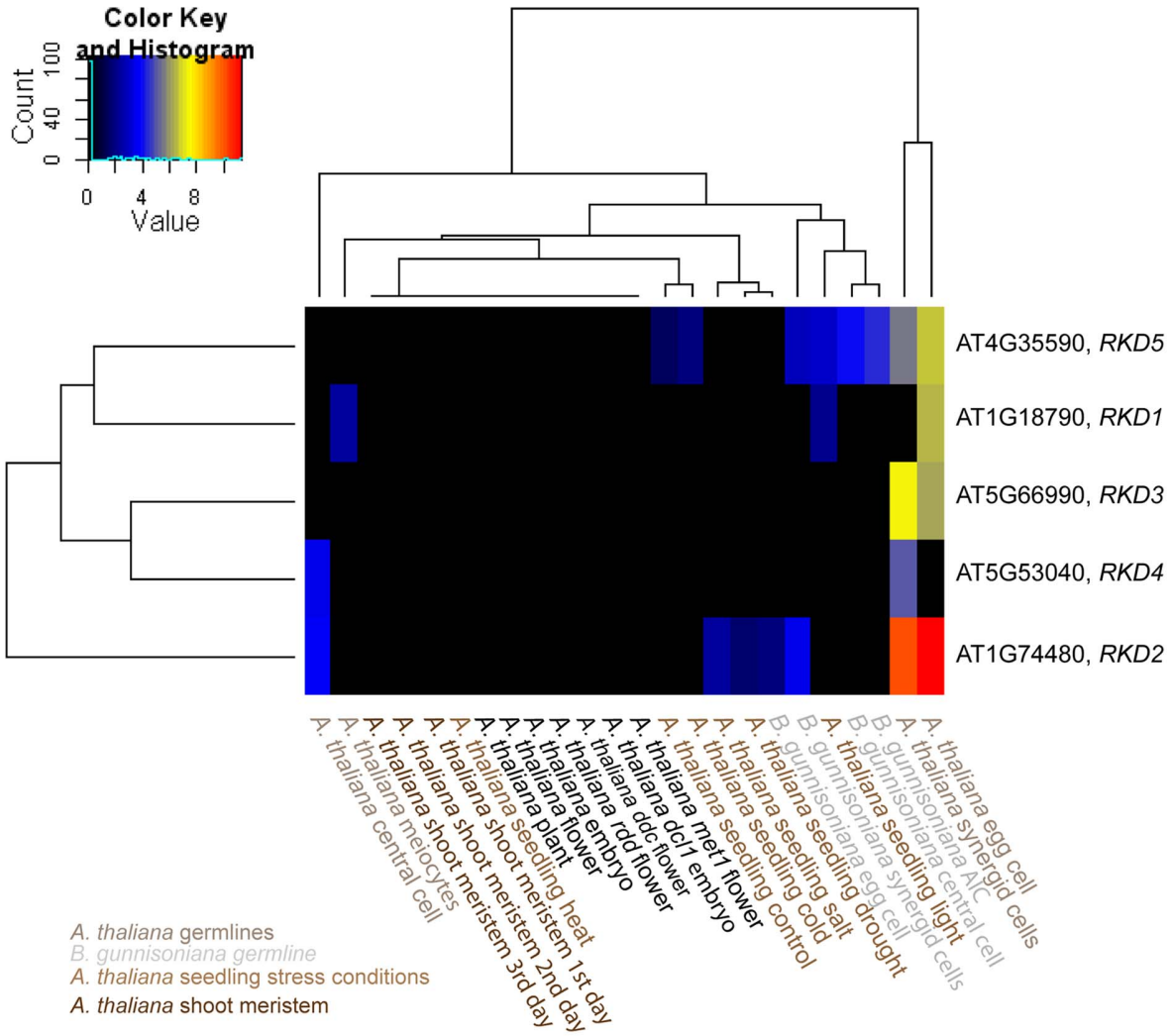


Figure 7. Heatmap clustering of members of the *AtRKD* gene family. Heatmap of normalized log2 transformed read counts from different *Arabidopsis* and *Boechera* cell- and tissue-types [13,62–66]. The hierarchical clustering of samples and genes was based on euclidean distance and hierarchical agglomerative clustering. No row scaling of colours was applied. Red denotes high expression and black low expression.
doi:10.1371/journal.pgen.1004476.g007

plant reproduction. Differences in activity were also observed for additional transcription factor families in agreement with previously identified differences in transcriptional regulation at later developmental stages in sexual and apomictic *P. simplex* flowers [35]. In the sexual *Arabidopsis* egg cell as compared to the apomictic *Boechera* egg cell, we observed the enriched expression of the RKD transcription factor family, which are important regulators of egg cell gene expression programs in *Arabidopsis* and wheat [67]. This suggests that RKD transcription factors might be specifically involved in the determination of the developmental fate of the sexual egg cell. Taken together, our findings indicate differences in the activity of important regulatory pathways during sexual and apomictic germline specification and development.

Germline specification during sexual reproduction and apomixis

Development of an unreduced embryo sac from an AIC is common to both diplospory and apospory. However, the founder cell of the female germline differs in position and cell fate between these two types of gametophytic apomixis. It is unknown whether a common regulator or a set of regulatory genes determines apomeiosis, or whether apomeiosis is mediated by unrelated developmental programs during apospory and diplospory. Interestingly, a number of important differences in gene expression have been observed in the aposporous initial cell in *H. praealtum* and the AIC of diplosporous *B. gunnisoniana*. This is consistent with the differences in cellular fate and identity between these apomicts. While the aposporous initial cell acquires a FMS-like fate without intervening cell division, the transcriptome of the AIC in a diplosporous apomict is expected to be more similar to the sexual MMC. This is in agreement with the lack of expression of several meiotic genes and other genes expressed in the sexual MMC in the aposporous initial cell in *H. praealtum* [31], differing from the transcriptome of the AIC in *Boechera*. Interestingly, in the *Boechera* AIC we did not observe evidence of expression of *DYAD/SWITCH*. In *Arabidopsis*, mutations in this gene have previously been shown to cause a diplospory-like phenotype with rare seed formation by the fertilization of unreduced egg cells [45].

The manipulation of cell cycle progression or meiotic genes has also been shown to lead to the formation of unreduced gametophytes [46,83, reviewed in 84]. The comparison between the *Arabidopsis* MMC and the *Boechera* AIC identified a number of core cell cycle genes to be differentially regulated. While a small number of genes important for meiotic cell cycle progression in *Arabidopsis* has already been described [46, reviewed in 84], detailed functional studies of candidate genes showing differential expression in the MMC and AIC will be required to elucidate their putative role in the discrimination between meiosis and apomeiosis. Interestingly, the *Arabidopsis* gene encoding WEE1 is only detected in the *Arabidopsis* MMC. The WEE1 protein is specifically removed to allow progression of mitosis [85]. In addition, homologues of three members of the *Arabidopsis* E2F transcription factor family have only been detected in the *Boechera* AIC but not in the *Arabidopsis* MMC. Members of this family are involved in the regulation of the centromer-specific histone 3 variant CENH3 in *Arabidopsis* [86]. Manipulation of CENH3 can induce genome elimination, a capacity that has already been successfully applied for the generation of synthetic clonal seeds from *Arabidopsis* in combination with *dyad* or MiMe mutants [83]. Based on our transcriptome analysis, different levels of *CENH3* expression have been observed in the *Boechera* germline as compared to *Arabidopsis*. In contrast to very low expression or absence in *Arabidopsis* gametes, higher expression levels of the *CENH3* homologue have been observed in *Boechera* gametes. It is thus possible that the absence of *DYAD/SWITCH* expression in the AIC combined with elevated expression levels of CENH3 in apomictic *Boechera* as compared to sexual gametes might play a role in naturally occurring diplospory. In addition to unknown parthenogenesis factors, the regulation of CENH3 activity might provide an additional control mechanism to secure the absence of a paternal contribution in the offspring.

While mutations in the gene encoding for *DYAD* lead to features of diplospory, mutations in *MEM*, *AGO9* and additional genes involved in a small RNA pathway have recently been reported to cause phenotypes reminiscent of apospory [12,42,43]. We identified additional genes involved in gene silencing and small RNA pathways to be differentially expressed in the MMC and the AIC. The expression of *ESP3* in the AIC is reminiscent of the previous identification of *ESP4* among the transcripts form the apospory-specific region in *P. squamulatum* [87]. This supports the importance of epigenetic regulatory pathways for sexual and apomictic reproduction.

Taken together, upon specification of the apomictic and sexual germline a number of differences involving regulatory processes such as hormone signalling, cell cycle control, and protein turnover have been observed. In addition, increased activity of signal transduction processes was identified as a typical feature of the apomictic germline. The potential role of positioning of the MMC or AIC and the signalling from the surrounding sporophytic tissues has previously been discussed [88,89], and our study has shown that signalling pathways are indeed modulated in the two modes of reproduction.

In conclusion, our study provides the first comprehensive transcriptional analysis of germline cells at key steps of apomictic reproduction in *B. gunnisoniana*. The generation and annotation of an apomictic reference transcriptome forms an essential basis for further analyses and allows the comparison of gene expression to *Arabidopsis* as sexual model species. Important differences in the development of the apomictic as compared to the sexual germline have been observed. While translational regulation is a feature conserved in both types of germline, polyamine and spermine/spermidine metabolism is only enriched upon initiation of the apomictic germline. In addition, key regulatory mechanisms are differentially regulated, involving hormone pathways, cell cycle control, signal transduction, and epigenetic regulatory processes. Thus, our analysis provides important new insights into gene regulation during apomictic germline development.

Methods

Plant material

A. thaliana Col-0 plants were used to isolate RNA for cloning of *in situ* probes. Plants were grown as described previously [12]. Seeds of *B. gunnisoniana* were kindly provided by Bitty Roy (University of Oregon, previously ETH Zürich) [48]. Seeds were surface sterilized and grown on MS plates for 10–14 days before transfer to a mixture of soil (ED73, Universalerde, Germany) and sand (5:1), fertilized with Plantomaag (Syngenta, Basel, Switzerland) and Osmocote (Scotts, Marysville, USA). Plants were grown in a greenhouse chamber with 60% humidity and 16 h light/ 8 h darkness at 20°C and 16°C, respectively.

Flow cytometry

Matured green seeds were harvested from *B. gunnisoniana* plants and individually analysed in a Quanta SC MPL flow cytometer (Beckman-Coulter, Nyon, Switzerland). Seeds were individually transferred to 1.2 ml cluster tubes (Thermo Scientific, Wohlen, Switzerland) containing 80 µl 0.1 M citric acid and 0.1% Triton X-100. A 3 mm stainless steel bead (Schieritz & Hauenstein AG, Zwingen, Switzerland) was added to each tube prior to shaking for 4 minutes at 30 Hz on a mixer mill (MM300, Retsch GmbH, Germany). Afterwards, 80 µl of 0.1 M citric acid containing 1% Triton X-100 was added and each tube was inverted 40 times. The solution containing the nuclei was filtered though fritted deep well plates (Nunc, Thermo Scientific, Wohlen, Switzerland) into 96-well V-bottom plates (Sarstedt, Numbrecht, Germany). Nuclei were collected by filtering in a centrifuge for 5 minutes at 150 g (Centrifuge 5810R, Eppendorf, Schönebuch, Switzerland). The nuclei were resuspended in 30 µl 0.1 M citric acid containing 1% Triton X-100. The samples were either analysed directly by flow cytometer robotics (Quanta SC MPL, Beckman-Coulter, Nyon, Switzerland) or stored at 4°C overnight prior to analysis. 120 µl of staining solution (0.4 M Na_2HPO_4, 2.6 ml H_2O, 27.4 µl DAPI (5.5 µg/µl), 0.2 µl β-mercaptoethanol) were added 2 min prior to analysis. The protocol was set to count nuclei for six minutes or until a maximum of 10'000 counts was reached. The Photo Multiplier Tube and the gain were set to have the embryo peak at around 200 on the linear fluorescent scale. *B. stricta* nuclei were used as external standard.

Cytological characterization

To quantitatively characterize developmental stages during germline development in *B. gunnisoniana*, material of 5 plants was used and averaged. Tissues were fixed in an ice-cold solution of ethanol:acetic acid (3:1; v/v), vacuum infiltrated on ice two times for 15 min, and left in fixative on ice over night before replacing the fixative with 70% ethanol. Tissues were cleared in chloral hydrate/glycerol/water (8:1:2; w/v/v), and microdissected with dissecting needles. Pictures were taken as previously described [12].

In situ hybridization

Genes for data confirmation by *in situ* hybridization were selected based on the following criteria: (1) expression in the *B.*

gunnisoniana AIC and no evidence of expression in the *A. thaliana* MMC, (2) representing different expression levels (Table S5), (3) high homology only to the respective homologue in *B. gunnisoniana* (82–96% identity between *A. thaliana* and *B. gunnisoniana* nucleotide sequences; Figure S3; Supporting Information S1), and (4) gene specificity in *A. thaliana*. Total RNA was isolated from *Arabidopsis* Col-0 inflorescences and from *B. gunnisoniana* buds and opened flowers using the RNeasy Plant Mini Kit (QIAGEN, Hilden, Germany). During the isolation procedure, RNA was treated with DNAseI on column. Reverse transcription was done as previously described ([12]; see Table S13 for a summary of primers and cDNA templates used). Fragment cloning and *in situ* hybridizations were done as previously described with modifications [12,51,90]: *in situ* hybridizations were performed on 8 µm thick sections of fixed and embedded *Boechera* buds or flowers. Pictures were taken and processed as previously described [12].

Laser-assisted microdissection

To prepare samples for LAM, buds with ovules harbouring the AIC were chosen as previously described for selection of buds with ovules harbouring the MMC in *Arabidopsis* [12] with modifications: for *Boechera* individual buds were harvested instead of inflorescences. To obtain ovules harbouring mature gametophytes, flowers were emasculated ~7 hours prior to fixation. The buds and flowers were fixed on ice in farmer's fixative (ethanol:acetic acid 3:1; v/v), vacuum infiltrated on ice two times for 15 min, and stored on ice over night before replacing the fixative with 70% ethanol. Embedding, microdissection, and LAM were done as previously described [12]. On average ~60 sections of AICs were collected per day, or ~25 sections for each cell-type of the mature female gametophyte. Egg and synergid cells from *Arabidopsis* were isolated as described previously for the central cell of *Arabidopsis* [13].

RNA isolation and quality control

LAM samples were stored dry at −80°C before RNA isolation. RNA isolation and quality control was done as previously described [12,13].

Array hybridization

RNA amplification and labelling was done with the MessageAmpII Kit (Ambion, Foster City, USA) as described previously. ~15 mg labeled aaRNA was fragmented and hybridized onto the *Arabidopsis* ATH1 GeneChip (Affymetrix) for 16 h at 45°C as described in the technical manual. The hybridization, staining, washing, and subsequent array scanning were performed as described previously [12]. Original data files are deposited under the Gene Expression Omnibus at NCBI (Accession Number GSE51996).

SOLiD sequencing

RNA isolation, amplification, library preparation, and SOLiD Sequencing were performed as described previously [12], except that SOLiD V4 was used for paired-end sequencing. Original data files are deposited in the NCBI database (Accession Number: SRP032961).

Reference transcriptome

As a tool for our data analysis we generated a reference transcriptome from female reproductive tissues of *B. gunnisoniana* at the two developmental stages of interest: (I) at megasporogenesis and (II) at the mature gametophyte stage. After isolation of mRNA and library preparation, sequencing was performed on an Illumina HiSeq 2000 instrument (see Supporting Methods S1 for details). Original data files are deposited in the NCBI SRA database (Accession Number SRP032960). The Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GBAD00000000. The version described in this paper is the first version, GBAD01000000.

Blast2GO annotation of the *B. gunnisoniana* reference transcriptome

After quality filtering, pre-processed reads were assembled using Trinity (version r2012-06-08) with default parameter settings, except that min_kmer_cov was set to 2. For annotation with Blast2GO, trinity assembled transcripts were compared to the NCBI non-redundant protein database (nr) using blastx (in blastall version 2.2.21). E-value cutoff was set to 0.00001. Top five hits were recorded. BLASTX results in XML format were analysed using b2g4pipe (version 2.5, [53]) to assign GO terms to the query transcript sequences.

BLAT comparison of the *B. gunnisoniana* reference transcriptome to TAIR10 cDNA

The BLAT (version 34) comparison of the *Boechera* reference transcriptome and the TAIR10 cDNA sequences (updated 12/14/2010) was done with default parameters for cross species DNA mapping (-q = dnax -t = dnax). The top hits were selected using the blat utility script pslCDnafilter (globalNearBest, globalNearBest plus minCov of 80%). TAIR10 cDNA annotation of the top hits was then transferred to the query transcripts.

Mapping of SOLiD reads

To obtain expression values based on the assembled *Boechera* reference transcriptome, short read data was processed as described in [13]. Gene-wise expression values were then defined as the sum of the expression values of individual transcript variants. Expression values based on the *A. thaliana* reference genome (TAIR10) were likewise calculated as described in [13].

Defining closest *Arabidopsis* homologues for *Boechera* gene models

To identify potential homologues of known genes from *A. thaliana* in the assembled reference transcriptome of *B. gunnisoniana* we used BLAT (version 34, [54]). Sequences from *Boechera* were aligned to *Arabidopsis* cDNAs (TAIR10), allowing for a maximal intron size (-maxIntron) of 2 kb. Individual alignment scores (bitScore) and lengths between a given pair of *Boechera* and *Arabidopsis* sequences were then summed up. For each gene of interest from *Arabidopsis*, the *Boechera* homologue was then defined as the gene with the highest bitScore sum (or none if no alignments were reported or the total alignment length was below 100 bp).

Analysis of sequence divergence

To estimate the extent of sequence divergence between a certain set of genes from *A. thaliana* and *B. gunnisoniana* we used ClustalX (version 2.1, [91]) with default settings (complete alignment, draw tree). Tree files were then used to cluster the genes in the heatmap plots (R packages ape, version 3.0-8 [92] and gplots, version 2.11.0, cran.r-project.org/web/packages/gplots/index.html).

*Bg*PANP

Microarray data were processed as described in [12], except using an updated annotation of the ATH1 microarray (brainar-

ray.mbni.med.umich.edu, TAIRG, version 14), and an alternative list of probesets for the background estimation ("negative probes"). Probe sequences were aligned to the assembled *Boechera* reference transcriptome using bowtie (version 0.12.7, [93]), allowing three mismatches. Probes without any alignments were considered as "negative probe" for the PANP algorithm [11].

Gene set enrichment studies using NOISeq and EdgeR

We used the NOISeq-sim algorithm (downloaded in April 2012, http://bioinfo.cipf.es/noiseq/doku.php, [57]) to analyse differential expression of genes between RNA-Seq samples of the *Boechera* germline (apo_initial3, egg_cell2, central_cell2, synergid_cell2). Reads were aligned to the *Boechera* reference transcriptome. The normalization method was set to tmm (Trimmed mean of M, [94]), no correction for feature length was applied, and default settings were used for all other parameters, including $q = 0.9$ as threshold to determine differentially expressed genes. Genes identified as significantly upregulated in all three pairwise comparisons of one cell type with the other three *Boechera* germline samples were described as enriched in the cell type. For higher stringency analysis EdgeR was used with the biological coefficient of variation (bcv) set to 0.8 and Benjamini-Hochberg multiple testing corrections. Genes with an adjusted p value (FDR) below 0.05 were considered to be significantly differentially expressed. To identify genes differentially expressed between egg cell and central cell we applied an unadjusted p value <0.001.

Identification of genes with evidence of expression only in *Boechera* or *Arabidopsis*

See Supporting Methods S1.

Mappings

Gene Ontology (GO) terms associated with *A. thaliana* genes were extracted from the functional descriptions and GOSLIM mappings available on TAIR (ftp.arabidopsis.org/home/tair/Proteins/TAIR10_functional_descriptions, and (ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt). GO terms associated with the genes of *Boechera* were obtained with b2g4pipe (version 2.5, [54]). Protein family (PFAM) and gene family (FAM) annotation was used as described [95].

GO, PFAM and FAM analyses

We used the Bioconductor package topGO [96] for gene ontology analysis. To test for overrepresentation of GO terms we used a Fisher's exact test in combination with the function "weight". As gene universe in the test for *Arabidopsis* MMC the whole ATH1 array genome was used, otherwise all genes annotated in the respective GO annotation were used. We used a two-sided Fisher's exact test and comparison against the gene universe as defined above to test for misrepresentation of protein family domains (PFAM) and gene families (FAM).

Heatmap clustering

Heatmaps were generated using the Bioconductor package gplots [97]. Hierachical agglomerative clustering (complete linkage) and euclidean distance was used. Normalization of RNA-Seq reads was done with the Bioconductor package DESeq [98]. Heatmaps were based on normalized log2-transformed total read counts for RNA-Seq data or log2-scale expression values generated by RMA for microarray data as previously described [12].

Venn diagrams

Venn diagrams were made with the online tool BioVenn (http://www.cmbi.ru.nl/cdd/biovenn/).

Supporting Information

Figure S1 Cytological characterization of reproductive development in *B. gunnisoniana*. (A, B) Development of the *B. gunnisoniana* AIC. A low percentage of AICs did not seem to divide (C) and likely arrested their development (D). (E) Dyad, (F) dyad with enlarged parietal cell or triad, (G) tetrad (artificially coloured in blue; based on the development of the integuments this tetrad is likely arrested or developmentally delayed), and (H) functional megaspore (FMS). (I, J) Mature gametophytes with unfused and fused polar nuclei, respectively. (K) Rarely more than one female gametophyte (artificially coloured in blue and pink) developed. (L–P) Seed development in young siliques after fertilization with embryo and endosperm development (L, M), young embryo developing in the absence of endosperm development (N), endosperm development without embryo development (O), and seed coat development in the absence of embryo or endosperm development (P). Black arrows point to AICs, stars mark (putative) parietal cells, white arrows point to dyads (or potential triads). Abbreviations: cen, central cell; egg, egg cell; syn, synergid cells; PN, polar nuclei; emb, embryo; end, endosperm. Scale bars are 40 μm. (Q) Summary of megaspore formation in *B. gunnisoniana*. In total 224 ovules were analysed. (R) Summary of mature gametophyte development in *B. gunnisoniana*. The percentages of mature gametophytes, gametophytes arrested at early developmental stages, gametophytes with an unexpected number of nuclei, and double gametophytes are given as analysed in 353 ovules.
(TIF)

Figure S2 Transcriptome analysis of the *Boechera* female gametes isolated by laser-assisted microdissection. (A) 6 μm thin section of a *Boechera* ovule harbouring the mature female gametophyte composed of egg cell, central cell, and synergid cells. Scale bar is 20 μm. (B, C) Venn diagrams showing the overlap of predicted expression in the *Boechera* and *Arabidopsis* female gametes. (B) Comparison of gene expression in the egg cell. Genes expressed in the *Arabidopsis* egg cell have been described before [11, reanalysed in 12] and were identified by RNA-Seq. Genes with evidence of expression in the *Boechera* egg cell were identified either by a P call with *Bg*PANP for the egg_cell1 sample or by at least 5 reads for homologues genes when mapped to the *Boechera* reference transcriptome. (C) Comparison of gene expression in the central cell. Genes expressed in the *Arabidopsis* central cell were previously identified using RNA-Seq [13]. Genes expression in the *Boechera* central cell was analysed by heterologous hybridization to the ATH1 microarray (central_cell1) or by RNA-Seq (central_cell2) by mapping the reads to the *Boechera* reference transcriptome and identification of *Arabidopsis* homologues.
(TIF)

Figure S3 Schematic alignment of *B. gunnisoniana* and *A. thaliana* genes selected for *in situ* hybridization. Schematic representation of gene exon and intron structures in *B. gunnisoniana* and *A. thaliana* for 5 genes selected for *in situ* hybridization. *Arabidopsis* gene and *Boechera* gene identifiers are given. The region selected for *in situ* probe design is indicated in red. Scaling is given in kb.
(TIF)

Figure S4 Heatmap of read counts for genes involved in silencing and small RNA pathways. Hierarchical clustering of log2 transformed read counts for 69 *Arabidopsis* genes homologues in *Boechera* involved

in small RNA and gene silencing pathways (as used in [12]). RNA-Seq data from different *Arabidopsis* and *B. gunnisoniana* cell- and tissue types were used [13,62–66]. The hierarchical clustering of samples and genes was based on euclidean distance and hierarchical agglomerative clustering. Colours are scaled per row. Red denotes high and black denotes low expression.
(TIF)

Figure S5 Heatmap of expression of *AGO* genes. (A) Hierarchical clustering of log2 transformed read counts of *AtAGO* genes and *Boechera* [13,62–66]. (B) Hierarchical clustering of log2 scale expression values of *AtAGO* genes in *Arabidopsis* as analysed by the RMA algorithm [12]. (A, B) The hierarchical clustering of samples and genes was based on euclidean distance and hierarchical agglomerative clustering. Colours were scaled by row. Red denotes high and black low expression.
(TIF)

Figure S6 Analysis of sequence divergence and heatmap of expression. Analysis of sequence divergence of members of the *ARI* (A) and *AGO* (B) gene family and read counts assigned.
(TIF)

Figure S7 Heatmap of expression of *ARI* genes. (A) Hierarchical clustering of log2 transformed read counts of *AtARI* genes and *Boechera* homologues including datasets from different transcriptional studies [13,62–66]. (B) Hierarchical clustering of log2 scale expression values of *AtARI* genes in *Arabidopsis* as analysed by the RMA algorithm [12]. The hierarchical clustering of samples and genes was based on euclidean distance and hierarchical agglomerative clustering. Colours were scaled by row. Red denotes high and black low expression.
(TIF)

Table S1 Annotation of *B. gunnisoniana* genes. *Boechera* genes were annotated using Blast2GO.
(ZIP)

Table S2 Assignment of *Arabidopsis* homologues to *Boechera* genes.
(TXT)

Table S3 *Bg*PANP expression calls. Datasheet with *Bg*PANP present (P) and absent (A) calls and p values in the AIC (apo_initial1, apo_initial2), the surrounding nucellus (sporo_nucellus1, and _2), egg cell (egg_cell1), central cell (central_cell), and synergid cells (synergid_cell) of *Boechera*.
(XLS)

Table S4 Expression values of individual variants of *Boechera* genes aligned to the reference transcriptome.
(TXT)

Table S5 Expression of genes selected for independent data confirmation by *in situ* analysis. P/A calls as analysed with *Bg*PANP for microarray samples and read counts for *B. gunnisoniana* homologues generated by mapping to the *B. gunnisoniana* reference transcriptome.
(PDF)

Table S6 Gene ontology (GO) analysis. (A) Molecular functions identified to be up-regulated based on 3'509 genes identified to be up-regulated in the egg cell as compared to the AIC, central cell and synergid cells. (B) Biological Processes identified to be up-regulated based on 1'806 genes identified to be up-regulated in the egg cell as compared to the AIC, central cell and synergid cells.
(PDF)

Table S7 Gene ontology (GO) analysis. Biological processes significantly upregulated in 142 genes enriched identified by EdgeR analysis in the *B. gunnisoniana* AIC as compared to the cell types of the mature female gametophyte.
(PDF)

Table S8 Analysis of protein family (PFAM) enrichment. Analysis of PFAM domains enriched in 852 genes with evidence of expression in the *Arabidopsis* MMC but not in the *B. gunnisoniana* AIC as tested by a two sided Fisher test. P values≤0.01 were considered significant.
(PDF)

Table S9 Analysis of protein family (PFAM) enrichment. Analysis of PFAM domains enriched in 2'146 genes with evidence of expression in the *Arabidopsis* but not in the *B. gunnisoniana* central cell as tested by a two sided Fisher test. P values≤0.01 were considered significant.
(PDF)

Table S10 Analysis of PFAM domains enriched in the apomictic *Boechera* germline. Significant enrichment of PFAM domains based on 901, 5'273 and 4'902 genes with evidence of expression in the AIC, egg cell and central cell of *Boechera* but not in the corresponding cell types of sexual *Arabidopsis* as analysed by two sided Fisher's exact test (p value<0.01).
(PDF)

Table S11 Enrichment of gene families in the *Boechera* female gametes. Significant enrichment of gene families based on 5'273 and 4'902 genes with evidence of expression in the egg cell and central cell of *Boechera* but not in the corresponding cell types of sexual *Arabidopsis* as analysed by two sided Fisher's exact test (p value< 0.01).
(PDF)

Table S12 Analysis of expression of core cell cycle genes. Lists of core cell cycle genes only found to be expressed in the AIC or the MMC, but not other way round.
(XLS)

Table S13 Primers and templates used for cloning of *in situ* probes.
(PDF)

Methods S1 Methods description on the generation of the *B. gunnisoniana* reference transcriptome. Methods description on the identification of genes with evidence of expression only in *Boechera* or in *Arabidopsis*.
(PDF)

Supporting Information S1 Alignment of *in situ* probe sequences from *Arabidopsis* to the homologues *B. gunnisoniana* genes generated by BLAT [54].
(ZIP)

Supporting Information S2 Description of genes involved in the small RNA pathway or in the DNA methylation pathway only detected in *Arabidopsis* or in *Boechera* and description on the influence of sequence similarities between *Arabidopsis* and *Boechera* homologues on the distribution of count data. As examples the *ARI* and the *AGO* gene families are discussed.
(PDF)

Acknowledgments

We are grateful to Samuel E. Wüst (University of Zürich) for helpful discussions, providing tools for data analysis, and for critical reading of the manuscript. We thank Catharine Aquino (Functional Genomics Center Zürich) for help with sequencing the reference transcriptome, Bitty Roy (ETH Zürich; presently University of Oregon) for providing *B. gunnisoniana* seeds and plants, Charles Spillane and Amal J. Johnston (University of Zürich; presently NUI Galway and University of Heidelberg,

respectively) for maintaining and propagating *B. gunnisoniana*, Valeria Gagliardini, Arturo Bolanos, Christof Eichenberger and Peter Kopf (University of Zürich) for general lab support, and Christian Frey and Karl Huwiler (University of Zürich) for plant care and greenhouse maintenance.

Author Contributions

Conceived and designed the experiments: AS UG. Performed the experiments: AS UCK CS DG MW. Analyzed the data: AS MWS UG UCK CS DG MW WQ. Contributed reagents/materials/analysis tools: WQ MWS AS PR UG. Wrote the paper: AS. Performed bioinformatic analyses: AS MWS WQ. Helped writing the paper: UG MWS. Participated in the design of RNA-Seq experiments and their analysis: PR. Revised and approved the manuscript: AS UG UCK CS DG MW MWS WQ.

References

1. Nogler GA (1984) Gametophytic apomixis. In: Embryology of Angiosperms. Edited by Johri BM. Berlin: Springer; pp. 475–518.
2. Savidan Y (2000) Apomixis: Genetics and breeding. Plant Breeding Reviews 18: 13–86.
3. Spillane C, Steimer A, Grossniklaus U (2001) Apomixis in agriculture: The quest for clonal seeds. Sex Plant Reprod 14: 179–87.
4. Bicknell RA, Koltunow AM (2004) Understanding apomixis: Recent advances and remaining conundrums. Plant Cell 16: 228–245.
5. Spillane C, Curtis MD, Grossniklaus U (2004) Apomixis technology development - virgin births in farmers' fields? Nat Biotechnol 22: 687–691.
6. Koltunow A, Grossniklaus U (2003) Apomixis: A developmental perspective. Annu Rev Plant Biol 54: 547–574.
7. Grossniklaus U (2001) From sexuality to apomixis: molecular and genetic approaches. In: The Flowering of Apomixis: From Mechanisms to Genetic Engineering. Edited by Savidan Y, Carman J, Dresselhaus T. Mexico DF: CIMMYT; pp.168–211.
8. Sprunck S, Gross-Hardt R (2011) Nuclear behavior, cell polarity, and cell specification in the female gametophyte. Sex Plant Reprod 24: 123–136.
9. Koltunow AM (1993). Apomixis: Embryo sacs and embryos formed without meiosis or fertilization in ovules. Plant Cell 5: 1425–1437.
10. Rodkienwicz B (1970) Callose in cell walls during megasporogenesis in angiosperms. Planta 93: 39–47.
11. Wüst SE, Vijverberg K, Schmidt A, Weiss M, Gheyselinck J, et al. (2010) *Arabidopsis* female gametophyte gene expression map reveals similarities between plant and animal gametes. Curr Biol 20: 506–512.
12. Schmidt A, Wüst SE, Vijverberg K, Baroux C, Kleen D, et al. (2011) Transcriptome analysis of the *Arabidopsis* megaspore mother cell uncovers the importance of RNA helicases for plant germline development. PLoS Biol 9: e1001155.
13. Schmid MW, Schmidt A, Klostermeier UC, Barann M, Rosenstiel P, et al. (2012) A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing. PLoS ONE 7: e29685.
14. Grossniklaus U, Nogler GA, van Dijk PJ (2001) How to avoid sex: the genetic control of gametophytic apomixis. Plant Cell 2001 13: 1491–1498.
15. Savidan Y (1982) Nature et hérédité de l' apomixie chez *Panicum maximum* Jacq. Trav. & Doc. Orstom 153: 1–159.
16. Nogler GA (1984) Genetics of apospory in apomictic *Ranunculus auricomus*. V Conclusion Bot Helv 94: 411–422.
17. Sherwood R T, Berg CC, Young BA (1994) Inheritance of apospory in buffelgrass. Crop Sci 34: 1490–1494.
18. Grimanelli D, Leblanc O, Espinosa E, Perotti E, González de León D, et al. (1998) Mapping diplosporous apomixis in tetraploid *Tripsacum*: one gene or several genes? Heredity 80: 33–39.
19. Ozias-Akins P, Roche D, Hanna WW (1998) Tight clustering and hemizygosity of apomixis-linked molecular markers in *Pennisetum squamulatum* implies genetic control of apospory by a divergent locus that may have no allelic form in sexual genotypes. Proc Natl Acad Sci U S A 95: 5127–5132.
20. Noyes RD, Rieseberg LH (2000) Two independent loci control agamospermy (apomixis) in the triploid flowering plant *Erigeron annuus*. Genetics 155: 379–390.
21. Valle CB, Miles JW (2001) Breeding of apomictic species. In: The Flowering of Apomixis: From Mechanisms to Genetic Engineering. Edited by Savidan Y, Carman J, Dresselhaus T. Mexico DF: CIMMYT; pp. 137–152.
22. Cáceres ME, Matzk F, Busti A, Pupilli F, Arcioni S (2001) Apomixis and sexuality in *Paspalum simplex*: characterization of the mode of reproduction in segregating progenies by different methods. Sex Plant Reprod 14: 201–206.
23. van Dijk PJ, Bakx-Schotman JM (2004) Formation of unreduced megaspores (diplospory) in apomictic dandelions (*Taraxacum officinale*, s.l.) is controlled by a sex-specific dominant locus. Genetics 166: 483–492.
24. Koltunow AM, Johnson SD, Rodrigues JC, Okada T, Hu Y, Tsuchiya T, et al. (2011) Sexual reproduction is the default mode in apomictic *Hieracium* subgenus *Pilosella*, in which two dominant loci function to enable apomixis. Plant J 66: 890–902.
25. Schranz ME, Kantama L, de Jong H, Mitchell-Olds T (2006) Asexual reproduction in a close relative of *Arabidopsis*: a genetic investigation of apomixis in *Boechera* (Brassicaceae). New Phytol 171: 425–438.
26. Grimanelli D, Leblanc O, Perotti E, Grossniklaus U (2001) Developmental genetics of gametophytic apomixis. Trends Genet 17: 597–604.
27. Sharbel TF, Voigt ML, Corral JM, Thiel T, Varshney A, et al. (2009) Molecular signatures of apomictic and sexual ovules in the *Boechera holboellii* complex. Plant J 104: 14026–14031.
28. Sharbel TF, Voigt ML, Corral JM, Galla G., Kumlehn J, et al. (2010) Apomictic and sexual ovules of *Boechera* display heterochronic global gene expression patterns. Plant Cell 22: 655–671.
29. Leblanc O, Armstead I, Pessino S, Ortiz JP, Evans C, et al. (1997) Non-radioactive mRNA fingerprinting to visualise gene expression in mature ovaries of *Brachiaria* hybrids derived from *B. brizantha*, an apomictic tropical forage. Plant Sci 126: 49–58.
30. Rodrigues JC, Carbral GB, Dusi DM, de Mello LV, Rigden DJ, et al. (2003) Identification of differentially expressed cDNA sequences in ovaries of sexual and apomictic plants of *Brachiaria brizantha*. Plant Mol Biol 53: 745–757.
31. Okada T, Hu Y, Tucker MR, Taylor JM, Johnson SD, et al. (2013) Enlarging cells initiating apomixis in *Hieracium praealtum* transition to an embryo sac program prior to entering mitosis. Plant Physiol 163: 216–31.
32. Vielle-Calzada JP, Nuccio ML, Budiman MA, Burson BL, Hussey MA, et al. (1996) Comparative gene expression in sexual and apomictic ovaries of *Pennisetum ciliare* (L.) Link. Plant Mol Biol 32: 1085–1092.
33. Sahu PP, Gupta S, Malaviya DR, Roy AK, Kaushal P, et al. (2012) Transcriptome analysis of differentially expressed genes during embryo sac development in apomeiotic non-parthenogenetic interspecific hybrid of *Pennisetum glaucum*. Mol Biotechnol 51: 262–271.
34. Pessino SC, Espinoza F, Martinez EJ, Ortiz JP, Valle EM, et al. (2001) Isolation of cDNA clones differentially expressed in flowers of apomictic and sexual *Paspalum notatum*. Hereditas 134: 35–42.
35. Polegri L, Calderini O, Arcioni S, Pupilli F (2010) Specific expression of apomixis-linked alleles revealed by comparative transcriptomic analysis of sexual and apomictic *Paspalum simplex* Morong flowers. J Exp Bot 61: 1869–83.
36. Ochogavía AC, Seijo JG, González AM, Podio M, Duarte Silveira E, et al. (2011) Characterization of retrotransposon sequences expressed in inflorescences of apomictic and sexual *Paspalum notatum* plants. Sex Plant Reprod 24: 231–246.
37. Barcaccia G, Varotto S, Meneghetti S, Albertini E, Porceddu A, et al. (2001) Analysis of gene expression during flowering in apomeiotic mutants of *Medicago* spp.: cloning ESTs and candidate genes for apomeiosis. Sex Plant Reprod 14: 233–238.
38. Chen L, Miyazaki C, Kojima A, Saito A, Adachi T (1999) Isolation and characterization of a gene expressed during early embryo sac development in apomictic Guinea grass (*Panicum maximum*) J Plant Physiol 154: 55–62.
39. Albertini E, Marconi G, Barcaccia G, Raggi L, Falcinelli M (2004) Isolation of candidate genes for apomixis in *Poa pratensis* L. Plant Mol Biol 56: 879–894.
40. Albertini E, Marconi G, Reale L, Barcaccia G, Porceddu A, et al. (2005) SERK and APOSTART. Candidate genes for apomixis in *Poa pratensis*. Plant Physiol 138: 2185–2199.
41. Tucker MR, Araujo AC, Paech NA, Hecht V, Schmidt ED, et al. (2003) Sexual and apomictic reproduction in *Hieracium* subgenus *pilosella* are closely interrelated developmental pathways. Plant Cell 15: 1524–1537.
42. Olmedo-Monfil V, Durán-Figueroa N, Arteaga-Vázquez M, Demesa-Arévalo E, Autran D, et al. (2010) Control of female gamete formation by a small RNA pathway in *Arabidopsis*. Nature 464: 628–632.
43. Garcia-Aguilar M, Michaud C, Leblanc O, Grimanelli D (2010) Inactivation of a DNA methylation pathway in maize reproductive organs results in apomixis-like phenotypes. Plant Cell 22: 3249–3267.
44. Singh M, Goel S, Meeley RB, Dantec C, Parrinello H, et al. (2011) Production of viable gametes without meiosis in maize deficient for an ARGONAUTE protein. Plant Cell 23: 443–458.
45. Ravi M, Marimuthu MP, Siddiqi I (2008) Gamete formation without meiosis in *Arabidopsis*. Nature 451: 1121–1124.
46. d'Erfurth I, Jolivet S, Froger N, Catrice O, Novatchkova M, et al. (2009) Turning meiosis into mitosis. PLoS Biol 7: e1000124.
47. Matzk F, Meister A, Schubert I (2010) An efficient screen for reproductive pathways using mature seeds of monocots and dicots. Plant J 21: 97–108.
48. Roy BA (1995) The breeding system of six species of *Arabis* (*Brassicaceae*). Am J Bot 82: 869–877.
49. Taskin KM, Turgut K, Scott RJ (2004) Apomictic development in *Arabis gunnisoniana*. Isr J Plant Sci 52: 155–160.
50. Aliyu OM, Schranz ME, Sharbel TF (2010) Quantitative variation for apomictic reproduction in the genus *Boechera* (*Brassicaceae*). Am J Bot 97: 1719–1731.

51. Johnston AJ (2007) Functional genomics of sexual and asexual reproduction in *Arabidopsis* and relatives. PhD Thesis, University of Zürich, Switzerland.
52. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnol 29: 644–654.
53. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.
54. Kent WJ (2002) BLAT-the BLAST-like alignment tool. Genome Res 12: 656–664.
55. Bar-Or C, Czosnek H, Koltai H (2007) Cross-species microarray hybridizations: a developing tool for studying species diversity. Trends Genet 23: 200–207.
56. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6: 377–382.
57. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: a matter of depth. Genome Res 21: 2213–2223.
58. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140.
59. Wijnker E, and Schnittger A (2013) Control of the meiotic cell division program in plants. Plant Reprod 26: 143–158.
60. Vandepoele K, Raes J, De Veylder L, Rouzé P, Rombauts S, et al. (2002) Genome-wide analysis of core cell cycle genes in *Arabidopsis*. Plant Cell 14: 903–916.
61. Gutierrez C (2009) The *Arabidopsis* cell division cycle. Arabidopsis Book 7: e0120.
62. Nodine MD, Bartel DP (2010) MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. Genes Dev 24: 2678–2692.
63. Chen C, Farmer AD, Langley RJ, Mudge J, Crow JA, et al. (2010) Meiosis-specific gene discovery in plants: RNA-Seq applied to isolated *Arabidopsis* male meiocytes. BMC Plant Biol 10: 280.
64. Torti S, Fornara F, Vincent C, Andrés F, Nordström K, et al. (2012) Analysis of the *Arabidopsis* shoot meristem transcriptome during floral transition identifies distinct regulatory patterns and a leucine-rich repeat protein that promotes flowering. Plant Cell 24: 444–462.
65. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Res 20: 45–58.
66. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BG, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133: 523–536.
67. Koszegi D, Johnston AJ, Rutten T, Czihal A, Altschmied L, et al. (2011) Members of the RKD transcription factor family induce an egg cell-like gene expression program. Plant J 67: 280–91.
68. Imai A, Matsuyama T, Hanzawa Y, Akiyama T, Tamaoki M, et al. (2004) Spermidine synthase genes are essential for survival of *Arabidopsis*. Plant Physiol 135: 1565–1573.
69. Wu XB, Wang J, Liu JH, Deng XX (2009) Involvement of polyamine biosynthesis in somatic embryogenesis of Valencia sweet orange (*Citrus sinensis*) induced by glycerol. J Plant Physiol 166: 52–62.
70. De-la-Peña C, Galaz-Avalos RM, Loyola-Vargas VM (2008) Possible role of light and polyamines in the onset of somatic embryogenesis of *Coffea canephora*. Mol Biotechnol 39: 215–24.
71. Dutra NT, Silveira V, de Azevedo IG, Gomes-Neto LR, Façanha AR, et al. (2013) Polyamines affect the cellular growth and structure of pro-embryogenic masses in *Araucaria angustifolia* embryogenic cultures through the modulation of proton pump activities and endogenous levels of polyamines. Physiol Plant 148: 121–32.
72. Ha HC, Sirisoma NS, Kuppusamy P, Zweier JL, Woster PM, et al. (1998) The natural polyamine spermidine functions directly as free radical scavenger. Proc Natl Acad Sci USA 95: 11140–11145.
73. Hörandl E, Hadacek F (2013) The oxidative damage initiation hypothesis for meiosis. Plant Reprod: DOI 10.1007/s00497-013-0234-7.
74. Martin MV, Fiol DF, Sundaresan V, Zabaleta EJ, Pagnussat GC (2013) *oiwa*, a female gametophytic mutant impaired in a mitochondrial manganese-superoxide dismutase, reveals crucial roles for reactive oxygen species during embryo sac development and fertilization in *Arabidopsis*. Plant Cell 25: 1573–1591.
75. Conner JA, Goel S, Gunawan G, Cordonnier-Pratt MM, Johnson VE, et al. (2008) Sequence analysis of bacterial artificial chromosome clones from the apospory-specific genomic region of *Pennisetum* and *Cenchrus*. Plant Physiol 147: 1396–1411.
76. Guilfoyle TJ, Ulmasov T, Hagen G (1998) The ARF family of transcription factors and their role in plant hormone-responsive transcription. Cell Mol Life Sci 54: 619–627.
77. Rademacher EH, Möller B, Lokerse AS, Llavata-Peris CI, van den Berg W, et al. (2011) A cellular expression map of the *Arabidopsis AUXIN RESPONSE FACTOR* gene family. Plant J 68: 597–606.
78. Craig KL, Tyers M (1999) The F-box: a new motif for ubiquitin dependent proteolysis in cell cycle regulation and signal transduction. Prog Biophys Mol Biol 72: 299–328.
79. Lechner E, Achard P, Vansiri A, Potuschak T, Genschik P (2006) F-box proteins everywhere. Curr Opin Plant Biol 9: 631–638.
80. Amiteye S, Corral JM, Vogel H, Sharbel TF (2011) Analysis of conserved microRNAs in floral tissues of sexual and apomictic *Boechera* species. BMC Genomics 12: 500.
81. Schallau A, Arzenton F, Johnston AJ, Hähnel U, Koszegi D, et al. (2010) Identification and genetic analysis of the *APOSPORY* locus in *Hypericum perforatum* L. Plant J 62: 773–784.
82. Mladek C, Guger K, Hauser MT (2003) Identification and characterization of the *ARIADNE* gene family in *Arabidopsis*. A group of putative E3 ligases. Plant Physiol 131: 27–40.
83. Marimuthu MP, Jolivet S, Ravi M, Pereira L, Davda JN, et al. (2011) Synthetic clonal reproduction through seeds. Science 331: 876.
84. Crismani W, Girard C, Mercier R (2013) Tinkering with meiosis. J Exp Bot 64: 55–65.
85. Cook GS, Grønlund AL, Siciliano I, Spadafora N, Amini M, et al. (2013) Plant WEE1 kinase is cell cycle regulated and removed at mitosis via the 26S proteasome machinery. J Exp Bot 64: 2093–2106.
86. Heckmann S, Lermontova I, Berckmans B, De Veylder L, Bäumlein H, et al. (2011) The E2F transcription factor family regulates CENH3 expression in *Arabidopsis thaliana*. Plant J 68: 646–656.
87. Zeng Y, Conner J, Ozias-Akins P (2011) Identification of ovule transcripts from the Apospory-Specific Genomic Region (ASGR)-carrier chromosome. BMC Genomics 12: 206.
88. Grossniklaus U, Schneitz K (1998) The molecular and genetic basis of ovule and megagametophyte development. Semin Cell Dev Biol 9: 227–238.
89. Tucker MR, Okada T, Johnson SD, Takaiwa F, Koltunow AM (2012) Sporophytic ovule tissues modulate the initiation and progression of apomixis in *Hieracium*. J Exp Bot 63: 3229–3241.
90. Vielle-Calzada J-P, Thomas J, Spillane C, Coluccio A, Hoeppner MA, Grossniklaus U (1999) Maintenance of genomic imprinting at the *Arabidopsis MEDEA* locus requires zygotic *DDM1* activity. Genes Dev 13: 2971–2982.
91. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.
92. Paradis E, Claude J, Strimmer K (2004) APE: Analysis of phylogenetics and evolution in the R language. Bioinformatics 20: 289–290.
93. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.
94. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11: R25.
95. Wüst SE, O'Maoileidigh DS, Rae L, Kwasniewska K, Raganelli A, Hanczaryk K, et al. (2012) Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. Proc Natl Acad Sci U S A 109: 13452–13457.
96. Alexa A, Rahnenführer J (2009) Gene set enrichment analysis with topGO. (www.bioconductor.org).
97. Warnes G, Bolker B, Lumley T (2010) gplots: Various R programming tools for plotting data. (www.bioconductor.org).
98. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106.

8.4 Selection-driven evolution of sex-biased genes is consistent with sexual selection in *Arabidopsis thaliana*

The following manuscript is published in "Molecular Biology and Evolution" (published by Oxford University Press, all rights reserved)[1]. I collected all microarray experiments, processed them, and performed the differential expression analyses (see first half of paragraph "Gene Expression Data Sets").

[1]Gossmann, TI, Schmid, MW, Grossniklaus, U, and Schmid, KJ (2014) Selection-driven evolution of sex-biased genes is consistent with sexual selection in *Arabidopsis thaliana*. Molecular Biology and Evolution 31: 574–583.

Selection-Driven Evolution of Sex-Biased Genes Is Consistent with Sexual Selection in *Arabidopsis thaliana*

Toni I. Gossmann,[1,2] Marc W. Schmid,[3] Ueli Grossniklaus,[3] and Karl J. Schmid*[,1]

[1]Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany

[2]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, United Kingdom

[3]Institute for Plant Biology and Zurich-Basel Plant Science Center, Plant Developmental Genetics, University of Zurich, Zurich, Switzerland

***Corresponding author:** E-mail: karl.schmid@uni-hohenheim.de.

Associate editor: John Parsch

Abstract

Sex-biased genes are genes with a preferential or specific expression in one sex and tend to show an accelerated rate of evolution in animals. Various hypotheses—which are not mutually exclusive—have been put forth to explain observed patterns of rapid evolution. One possible explanation is positive selection, but this has been shown only in few animal species and mostly for male-specific genes. Here, we present a large-scale study that investigates evolutionary patterns of sex-biased genes in the predominantly self-fertilizing plant *Arabidopsis thaliana*. Unlike most animal species, *A. thaliana* does not possess sex chromosomes, its flowers develop both male and female sexual organs, and it is characterized by low outcrossing rates. Using cell-specific gene expression data, we identified genes whose expression is enriched in comparison with all other tissues in the male and female gametes (sperm, egg, and central cell), as well as in synergids, pollen, and pollen tubes, which also play an important role in reproduction. Genes specifically expressed in gametes and synergids show higher rates of protein evolution compared with the genome-wide average and no evidence for positive selection. In contrast, pollen- and pollen tube-specific genes not only have lower rates of protein evolution but also exhibit a higher proportion of adaptive amino acid substitutions. We show that this is the result of increased levels of purifying and positive selection among genes with pollen- and pollen tube-specific expression. The increased proportion of adaptive substitutions cannot be explained by the fact that pollen- and pollen tube-expressed genes are enriched in segmental duplications, are on average older, or have a larger effective population size. Our observations are consistent with prezygotic sexual selection as a result of interactions during pollination and pollen tube growth such as pollen tube competition.

Key words: **adaptive evolution, angiosperms, reproduction, pollen competition.**

Article

Introduction

The role of different evolutionary forces in the evolution of reproductive genes is a central question in population genetics and molecular evolution (Parsch and Ellegren 2013). Sex-linked and sex-biased genes, the latter showing preferential expression in one of the sexes, tend to show accelerated rates of evolution in animal species (Mank et al. 2007; Meisel 2011; Grath and Parsch 2012), which in some cases results from positive selection (Clark and Swanson 2005; Pröschel et al. 2006; Baines et al. 2008). Several hypotheses were proposed to explain the variation in the level of selective forces between classes of genes accounting for sexual differences. First, the faster-X hypothesis predicts that X-linked loci show higher rates of adaptive evolution compared with autosomal genes assuming that beneficial mutations are, on average, recessive (Charlesworth et al. 1987). This effect is particularly pronounced for male-biased X-linked genes (Baines et al. 2008; Mank et al. 2010). Another factor influencing rates of evolution is the effective population size (N_e) because it is a major determinant of selection efficiency (Vicoso and Charlesworth 2009; Gossmann et al. 2012). As N_e varies across the genome and between autosomal and sex chromosomes (Mank et al. 2010; Gossmann et al. 2011), a smaller N_e in nonrecombining sex chromosomes may contribute to rapid evolution. A third process is sexual antagonism, in which genes are advantageous for one, but disadvantageous for the other sex. This leads to genomic conflicts, which may be mediated or resolved by gene duplication, sex-specific gene expression, or parent-of-origin-dependent expression, that is, genomic imprinting (Rice 1996; Spillane et al. 2007). Finally, sexual selection occurs within species via differential mate choice, sperm competition, and sperm–egg recognition (Bernasconi et al. 2004). The latter two are assumed to be major determinants for increased evolutionary rates in male genes in humans, *Drosophila melanogaster*, and mice (Price et al. 1999; Wyckoff et al. 2000; Dorus et al. 2010).

Compared with animals, much less is known about the molecular evolution of reproductive genes in flowering plants. Sex-chromosomes are found only in few genera like *Silene*, *Papaya*, or *Asparagus* and they are of recent evolutionary origin (from 0.5–2.2 Ma for *Papaya* to 8–24 Ma for *Silene* (Charlesworth 2002; Ming et al. 2011; Gschwend et al. 2012). *Arabidopsis thaliana*, like the majority of plant species ($\approx$72%; Yampolsky and Yampolsky 1922; Dellaporta and Calderon-Urrea 1993), is a hermaphrodite, whose flowers harbor male and female reproductive organs. It is highly

homozygous due to its mode of reproduction by self-fertilization and shows little evidence for adaptive evolution on a genome-wide scale (Bustamante et al. 2002; Gossmann et al. 2010; Cao et al. 2011; Slotte et al. 2011). Initially, the lack of evidence of positive selection was explained by the high rate of inbreeding (Bustamante et al. 2002), but comparisons with the outcrossing relative *A. lyrata* have not confirmed this hypothesis (Foxe et al. 2008; Gossmann et al. 2010).

The morphological development of reproductive organs has been studied in great detail in flowering plants (Dickinson and Grant-Downton 2009). Unlike in animals, the reproductive cells are not set aside early in development but produced during flower development from specific sporophytic cells, the archespores (Ma and Sundaresan 2010; Grossniklaus 2011; Schmidt et al. 2012). These eventually undergo meiosis to form haploid spores that develop into multicellular male and female gametophytes. Later during gametophyte development, more specialized cells are formed, including sperm and egg cells (Dickinson and Grant-Downton 2009; Twell 2011). After pollination, complex pollen–pistil regulatory mechanisms initiate or inhibit the growth of the pollen tube into the pistil (Chapman and Goring 2010; Kessler and Grossniklaus 2011) and later into the ovule (Dresselhaus and Sprunck 2012). A pair of immotile sperm cells is transported to the female gametes via the pollen tube to allow double fertilization of one egg cell and one central cell, which will form the embryo and endosperm, respectively (Palanivelu and Tsukamoto 2011; Dresselhaus and Sprunck 2012).

Knowledge about evolutionary processes associated with genes expressed during the process of pollination, pollen tube growth, and fertilization is still limited in plants, even though the molecular mechanisms are being investigated in detail. To close this gap, we used resequencing data of 80 *A. thaliana* accessions (Cao et al. 2011) and the genome of *A. lyrata* as an outgroup (Hu et al. 2011) to infer how purifying and positive selection act on genes specifically expressed in reproductive tissues. We focus on male and female gametophytes that consist of the gametes (sperm, egg, and central cell) and accessory cells essential for fertilization (vegetative cell growing the pollen tube and synergids) and compare our results in the light of evolutionary hypotheses developed for sex-biased genes derived from animal species. They include predictions about rates of molecular protein evolution for genes specifically expressed in one sex, the roles of gene duplication, intraspecific sexual conflicts, differences in N_e, and the strength of selection. The ratio of nonsynonymous to synonymous per site substitution rates (d_n/d_s) is used to infer predominant types of selection acting on subsets of genes, and the distribution of fitness effects (DFE) of new mutations as well as a derivative of the McDonald–Kreitman test (McDonald and Kreitman 1991) are used to estimate the proportion of adaptive substitutions (Eyre-Walker and Keightley 2009).

We identified significant differences between male- and female-specific genes. In particular, we show that male genes specifically expressed in the pollen and pollen tube have lower rates of protein evolution, which likely is the consequence of stronger purifying selection acting on new mutations. However, even though the rate of protein evolution is reduced for those genes, the contrast of intraspecific diversity and interspecific divergence suggests that up to 30% of nonsynonymous changes were caused by positive selection. These differences in selective pressure may result from pollen–pollen interactions and competition during pollen tube growth, and therefore suggest the presence of sexual selection during *A. thaliana* pollination (Moore and Pannell 2011).

Results

Male and Female Gametophytes Differ in Their Number of Specifically Expressed Genes

In this study, we focused on genes expressed in the haploid male and female gametophytes and did not consider sporophytic reproductive tissues. Specifically, we used publicly available expression data from male (sperm) and female gametes (egg and central cell), and from two accessory cell types with a central role in reproduction. They are the vegetative cell of the pollen which germinates and grows a pollen tube that transports the sperm cells, and the synergids of the female gametophyte which attract the pollen tube and interact with it during the fertilization process. The microarray design included 21,428 annotated genes and the number of hybridization experiments varied between tissues (three hybridization experiments for central, synergid, egg, and sperm cells, respectively, 13 for pollen tubes and 23 for pollen). Counts of significantly expressed genes (at least two hybridization signals with $P < 0.05$) were 9,213 for egg cells, 9,259 for central cells, and 7,534 for synergides. In male tissues, 14,159 genes were significantly expressed in pollen, 11,657 genes in pollen tubes, and 7,832 genes in sperm cells.

We identified genes that are specifically enriched in each of the six tissues. Taken together, 1,019 and 196 genes were enriched in the male and female gametophytes, respectively. Supplementary figure S1 (Supplementary Material online) shows the overlaps between the gene sets identified in the comparisons. In the female gametophyte, only about one-fourth of genes were shared by at least two cell types, whereas the rest shows cell type-specific expression (supplementary fig. S1*a*, Supplementary Material online). Genes enriched in the male gametophyte cluster into three major groups, which are specific to sperm cells (399 genes), pollen and pollen tubes (415 genes), or pollen tubes (111 genes, supplementary fig. S1*b*, Supplementary Material online). In subsequent analyses, we used either all the genes enriched in the gametophytes as male and female gametophytic standards (termed male and female genes, respectively) or the specific gene sets (egg cell, central cell, synergids, sperm, pollen, pollen tube; supplementary data set S1, Supplementary Material online). We also identified 20 genes that were expressed specifically in both male and female gametophytes.

Male Genes Are Enriched in Segmentally Duplicated Blocks

To functionally annotate specifically expressed genes in the male and female gametophytes, we performed a gene ontology (GO) term enrichment analysis. Among the top

GO terms for male tissues were "developmental cell growth," "pollen tube growth," and "enzyme regulator activity" (supplementary data set S1, Supplementary Material online). Altogether, 115 GO terms were overrepresented among male genes. Only 11 terms were significantly enriched among female genes, including "endomembrane system" and "pectinesterase enzymatic activity." We investigated the physical location of reproductive genes in the genome and found that male and female genes are homogeneously distributed across the five chromosomes. However, on a smaller genomic scale (10 Mb) male genes show a nonrandom distribution (χ^2 test, $P = 9 \times 10^{-3}$). Genes enriched in synergid and sperm cells show a nonrandom distribution across the genome on a 10 Mb genomic scale (χ^2 test, $P = 0.04$ and $P = 0.02$, respectively). To investigate this pattern in greater detail, we tested whether reproductive genes occur more or less often in duplicated blocks than expected by chance. Such segmental blocks mainly originated from two whole-genome duplication (WGD) events 20–40 Ma (Jiao et al. 2012). Male genes occur more often in duplicated blocks than expected by chance (χ^2 test, $P = 2.1 \times 10^{-3}$), which is caused by genes expressed in pollen and pollen tubes (χ^2 test, $P = 4.2 \times 10^{-3}$ and $P = 2.1 \times 10^{-4}$, respectively), but not in sperm cells (χ^2 test, $P = 0.36$). We determined the extent of tissue specificity in gene expression using the tissue specificity index τ (see Materials and Methods). Male and female genes do differ with respect to their average τ values with female genes being more specifically expressed than male genes (t–test, $P = 1.0 \times 10^{-15}$). Furthermore, there was no evidence that male genes from duplicated blocks tend to be more specifically expressed when compared with male genes that are located outside duplicated blocks (t–test, $P = 0.47$). Gene duplications are thought to be a major reason for genetic novelties caused by neo- or subfunctionalization of one or both of the duplicated gene copies. Tissue-specific expression is a major indicator for sub- or neofunctionalization where complementary expression patterns are considered as evidence for subfunctionalization (Liu et al. 2011). As nearly one-half of the male genes occur in duplicated blocks (496 out of 1,019), we examined the tissue specificity of these genes in comparison with their corresponding paralogs. Expression of male-specific genes showed greater tissue specificity (U test, $P = 2.2 \times 10^{-11}$), although the expression of the corresponding paralogs is more specific in comparison with random genes from duplicated blocks (U test, $P = 1.3 \times 10^{-14}$). However, we observed a bimodal distribution in the specificity of expression (fig. 1) because among 75 paralog pairs ($\approx$30% of the 496 genes) both copies show specific expression in male tissues. Only six of those paralogous pairs are expressed in different male tissues, and in three gene pairs the corresponding paralog of a male gene is enriched in a female tissue (supplementary table S1, Supplementary Material online).

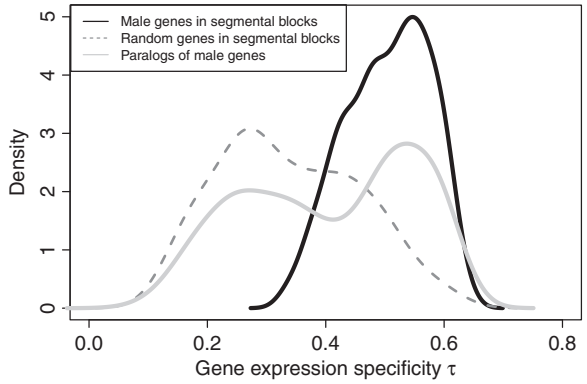


FIG. 1. Expression specificity τ (Yanai et al. 2005) for genes from segmental blocks. The average tissue specificity of the corresponding paralogs (solid gray) is reduced compared with the male gene sets (solid black) but increased in comparison with random genes from duplicated blocks (dashed). The discrete distributions were smoothed using kernel density estimates.

Male and Female Genes Differ in Their Evolutionary Age and Their Long-Term Evolutionary Rates

Using sequence data from *A. lyrata*, we estimated the interspecific sequence divergence for male, female, and random genes. The random genes consist of 500 genes sampled without replacement from the *A. thaliana* reference genome excluding reproductive genes, resulting in 476 random genes. The rate of protein evolution since the split was estimated from the ratio of nonsynonymous to synonymous per site substitutions ($\omega = d_n/d_s$). Protein divergence was higher for female genes (median $\omega = 0.29$) compared with male genes or a random set of genes (median $\omega = 0.17$ and $\omega = 0.16$; U test, $P = 3.3 \times 10^{-13}$ and 3.5×10^{-9}, respectively). There is no difference between male and random genes (U test, $P = 0.07$). Higher ω ratios observed in female genes were caused by differences in d_n values (U test, $P = 1.1 \times 10^{-16}$ and 2.1×10^{-10}, female vs. male and random, respectively), but not in d_s values (U test, $P = 0.07$ and $P = 0.08$). As gene age is a strong predictor of the rate of molecular evolution in primates (Cai and Petrov 2010), we estimated the gene age by the most distant pairwise best Blast hit in a set of ten genomes arranged in a hierarchical order representing their phylogenetic relationship (supplementary fig. S2, Supplementary Material online). Female and male genes show distinct age distributions (fig. 2, χ^2 test, $P = 2.3 \times 10^{-7}$), with female genes being significantly younger than male genes (χ^2 test, $P = 1.4 \times 10^{-9}$). Gene age estimates that are based on single best Blast hits (Quint et al. 2012) produced very similar results (supplementary fig. S3, Supplementary Material online, χ^2 test, $P = 1.8 \times 10^{-7}$). Therefore, the distinct age distribution is independent of the method gene age was estimated, even though numerous WGD events occurred during angiosperm evolution, which may severely bias the outcome of gene age estimates depending on the applied method.

Higher Levels of Purifying Selection and Higher Proportions of Adaptive Substitutions in Pollen and Pollen Tube Genes

To estimate whether differential evolution is reflected in polymorphism levels, we estimated the proportion of

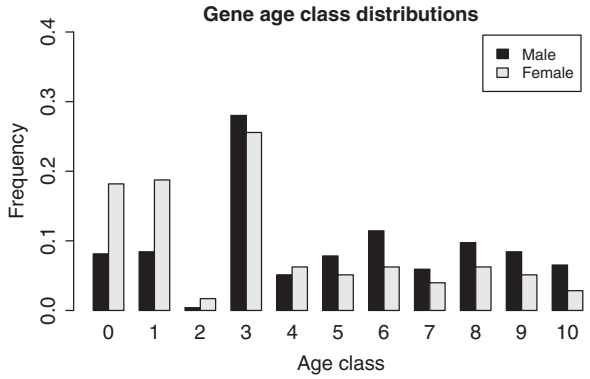


Fig. 2. Distribution of male and female reproductive genes regarding their gene age.

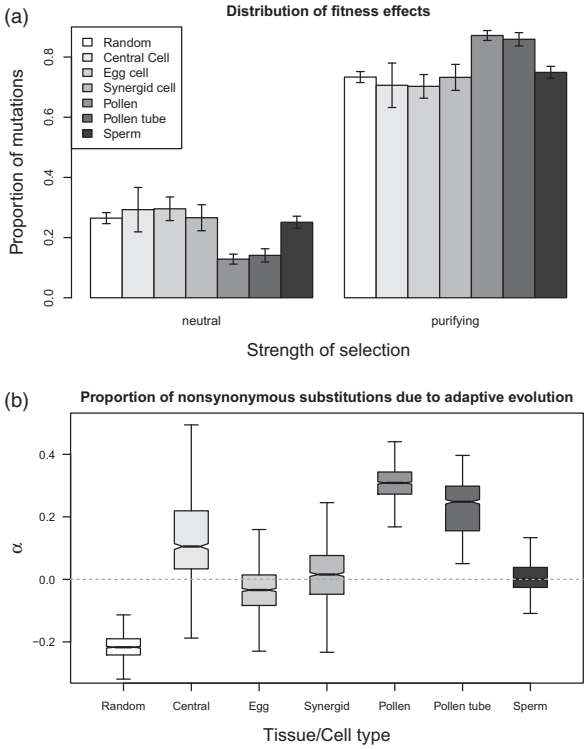


Fig. 3. Estimates of purifying and positive selection in *Arabidopsis thaliana* for random genes and genes with enriched expression in six reproductive tissues. (*a*) Estimated proportions of nonsynoynmous mutations with effectively neutral and purifying selection effects. (*b*) Estimates of the proportion of nonsynonymous substitutions due to adaptation (α).

mutations that are effectively neutral (f) and deleterious ($1-f$) for six reproductive tissues and random genes (fig. 3*a*). Among the female sets and genes enriched in sperm cells, $1-f$ is very similar and ranges from 70% to 72%. In contrast, values for genes expressed in pollen tubes and pollen were higher (86–87%), indicating that in the latter two groups a larger fraction of nonsynonymous mutations evolved under purifying selection. The inference of selective effects from polymorphism frequencies can be used to obtain an improved estimate of the proportion of nonsynonymous fixations resulting from positive selection, α (Eyre-Walker and Keightley 2009). Among the six reproductive gene sets, only pollen- and pollen tube-enriched genes show α values that are significantly greater than zero (fig. 3*b*). The rate of adaptive substitutions relative to the rate of synonymous substitutions, ω_a, of all reproductive genes was larger than genome-wide estimates, although the difference is only significant for the male gene sets (supplementary fig. S4, Supplementary Material online). However, the ω_a ratio for genes expressed in central cells was similar to estimates for pollen tube and pollen genes, suggesting that positive selection may act on female reproductive genes, even though this effect was not significant. There were no significant differences between genes from the duplicated blocks or for genes that are *Brassicacae*-specific (gene age < 3; fig. 2). Furthermore, the nucleotide diversity for sperm cell-specific genes is significantly reduced (vs. random genes, U test, $P = 0.03$) and those genes are located in genomic regions with reduced population recombination rates (ρ, Horton et al. [2012], random genes vs. sperm genes, U test, $P = 0.02$). The other gene sets did not differ from the random gene set with respect to their nucleotide diversity and recombination rates. As the 80 *A. thaliana* accessions originated from eight distinct geographic regions, we obtained estimates of $1-f$ for each of the eight regions (supplementary fig. S5, Supplementary Material online) to exclude that the observed differences are an artifact of the population structure. In all populations, the proportion of deleterious mutations was highest for pollen and pollen tube genes and lowest for one of the female sets. However, populations differ with respect to the $1-f$ estimates, which may correlate with variation in N_e among populations (Cao et al. 2011). In fact, the proportion of deleterious mutations ($1-f$) was positively correlated with N_e in the male sets, but not in the female sets or among random genes (supplementary table S2, Supplementary Material online). We also estimated the rates of adaptive evolution using a different sampling scheme in which we randomly sampled one allele per population and found that estimates of f, α, and ω_a are very similar to the initial sampling scheme, which did not take population structure into account (supplementary fig. S6, Supplementary Material online).

Discussion

Expression Patterns and Genomic Location

On the basis of microarray gene expression data, we identified sets of genes, which show significantly enriched expression in reproductive tissues in *A. thaliana*. We differentiated between three male (sperm, pollen, and pollen tube) and three female (synergids, egg cell, and central cell) tissues or cell types, and focused on genes whose expression was significantly enriched in at least one reproductive tissue or cell type. The male set consists of roughly five times more genes than the female set. The difference in the number of specific genes is partly the result of the focus of expression studies on male plant reproductive tissues,

which is experimentally easier to dissect and isolate. Only few studies investigated plant female cell-specific expression in greater detail (Wuest et al. 2010; Schmidt et al. 2011; Schmid et al. 2012). Consequently, the strong contrast between the number of overrepresented GO terms in male and female tissues may partly be caused by the lack of statistical power for the female set and an experimental bias because male tissues have been investigated to a greater extent (Schmidt et al. 2012). Additionally, there is a dependency on the design of the microarrays with relevant genes being absent from the chip (Schmidt et al. 2012). Future studies using RNAseq will provide further insights and circumvent the shortcomings of microarray-based expression analyses (Schmidt et al. 2012).

There were significant differences between male and female genes with respect to their genomic distribution. Genes with enriched expression in pollen and pollen tubes are more frequently located in segmentally duplicated blocks that originated from WGD events. This overrepresentation is caused by the corresponding paralogs showing specific expression in the same reproductive tissue, supporting the notion that segmentally duplicated genes show a strong correlation of coexpression across tissues (Gossmann and Schmid 2011). Pollen genes are enriched in segmentally duplicated blocks but not among tandem duplicated genes (Liu et al. 2011). In *Caenorhabditis elegans* and *D. melanogaster*, male-biased genes have a significantly increased proportion of paralogs (Cutter and Ward 2005; Gnad and Parsch 2006), which is a consequence of subfunctionalization (Ellegren and Parsch 2007). This is also a likely scenario in *A. thaliana*, but in contrast to the two animal species, the higher proportion of male-biased paralogs is a consequence of WGD events and not of tandem duplications. A previous study (Gossmann and Schmid 2011) identified 14 pairs of paralogs with a different rate of molecular evolution between the two paralogs (supplementary table S3, Supplementary Material online) and specific expression of at least one paralog in a reproductive tissue. Interestingly, none of these paralog pairs are specifically expressed in the same tissue. We observed only few examples of paralogous genes that are expressed in different reproductive tissues, suggesting a limited impact of duplications in resolving sexual conflicts in *A. thaliana*.

Evolutionary and Adaptive Rates Differ among Reproductive Tissues

Estimates of the proportion of advantageous nonsynonymous substitutions, α (Fay et al. 2001; Smith and Eyre-Walker 2002), and the proportion of deleterious mutations, $1-f$ (Slotte et al. 2011), reveal the long-term influence of positive selection and the short-term effect of selection against deleterious mutations in a population. Per-chromosome estimates for *A. thaliana* based on a different data set showed that the proportion of nonsynonymous polymorphisms with evidence of purifying selection is $1-f = 73-77\%$ (Slotte et al. 2011), which is similar to our estimate of 73% for the random gene set excluding reproductive genes (fig. 3*a*).

Genes expressed in pollen and pollen tubes of *A. thaliana* show higher α values and a higher proportion of polymorphisms under purifying selection than random genes. In contrast, genes with enriched expression in female cell types and sperm cells show higher rates of protein evolution, which is likely the consequence of relaxed purifying selection on segregating variation and not due to increased α values. Generally, positive selection in protein coding genes does not seem to be a strong evolutionary force in species of the genus *Arabidopsis* (i.e., $\alpha \approx 0$; Foxe et al. 2008; Gossmann et al. 2010; Slotte et al. 2011), although it was detected in some functional groups such as disease resistance and abiotic stress tolerance genes (Slotte et al. 2011). Signatures of adaptive evolution in genes expressed in pollen and pollen tubes are similar to increased rates of adaptive evolution in male-biased genes of *D. melanogaster* (Pröschel et al. 2006; Baines et al. 2008). However, genes expressed specifically in reproductive tissues of *D. melanogaster* and *D. pseudoobscura* show the highest rates of protein evolution, but no higher rates of adaptation (Meisel 2011; Grath and Parsch 2012). There is also limited evidence for higher rates of protein evolution but not for higher rates of positive selection in *D. ananassae* (Müller et al. 2012). Varying levels of protein evolution can be the consequence of different levels of purifying or positive selection. Genome-wide levels of adaptive substitutions are high in *Drosophila* potentially due to its large effective population size of about 800,000 for *D. melanogaster* and *D. pseudoobscura* (Gossmann et al. 2012). *Arabidopsis thaliana* has a relatively low effective population size of $\approx$275,000 (Cao et al. 2011), and its genome evolves predominantly under purifying selection (Slotte et al. 2011). Therefore, only mutations with fairly strong effects will be effectively selected.

It has been proposed that the higher mutation rate observed in males is due to differences in the number of cell replications during male and female gamete development (Hurst and Ellegren 1998; Ellegren 2007). Such a bias in cell divisions also seems to occur in dioecious plants (Filatov and Charlesworth 2002) and a possible explanation for a male-biased transmission of mutations in *A. thaliana* has been described (Whittle and Johnston 2002, 2003). If the rate of adaptation is limited by the supply of adaptive mutations, one expects that male-biased genes show higher rates of adaptation because they are mainly expressed in the haploid state where mutations are immediately exposed to selection. According to this model, genes specifically expressed in sperm, pollen, and pollen tubes should be affected in a similar fashion. Three explanations may account for the observed differences between pollen/pollen tube and sperm genes. First, nucleotide diversity at synonymous sites θ_S is significantly reduced in sperm cell-specific genes (supplementary fig. S7*a*, Supplementary Material online). There is evidence that N_e varies within the genome of *A. thaliana* (Gossmann et al. 2011), which may be the consequence of variation in recombination rate, selective pressures, and Hill–Robertson effect (Hill and Robertson 1966). If mutation rates do not differ as suggested by the synonymous divergence estimates (U test, $P = 0.26$), it is possible that they have a smaller N_e

and, therefore, selection is less efficient. Second, genes expressed in sperm cells are located in genome regions with lower population recombination rates, for which we found evidence (supplementary fig. S7*b*, Supplementary Material online; Horton et al. 2012). This may be a consequence of a reduced N_e or lower rates of recombination per generation, both of which result in a reduced efficacy of selection (Gossmann et al. 2011). Third, expression levels are negatively correlated with d_n/d_s and d_n in *Arabidopsis* and *Medicago* (Slotte et al. 2011; Yang and Gaut 2011; Paape et al. 2013), and expression is indeed lowest for genes expressed in sperm cells among all reproductive tissues, and is significantly different from pollen and pollen tube genes (supplementary fig. S7*c*, Supplementary Material online, *U* test, sperm only genes vs. pollen only genes, $P = 7.5 \times 10^{-20}$). Using a generalized linear model we find that expression intensity is correlated to d_n/d_s after controlling for variation in τ ($P = 8.1 \times 10^{-4}$), which appears not to be the case in the other reproductive tissues. Taken together, this implies that mutations specifically expressed in sperm cells evolve under more relaxed selection than pollen and pollen tube-specific genes.

In contrast to *Drosophila* or mammals, sperm cells are nonmotile in *Arabidopsis*, suggesting that interactions between pollen grains during pollination and pollen tube growth are plausible targets of prezygotic sexual selection (Carlson et al. 2009, 2011). In *A. thaliana*, 60% of pollen tubes grow to the four nearest ovules (Hülskamp et al. 1995), indicating that pollen tube growth speed is a crucial factor for successful fertilization (Williams 2008). Pollen tube growth in *A. thaliana* is influenced by sporophytic cells of the pistil and then guided and received by the female gametophyte (Kessler and Grossniklaus 2011; Takeuchi and Higashiyama 2011). In this study, a sporophytic tissue–pollen interaction would not be reflected among female genes but in the pollen and pollen tube gene set. Therefore, if interactions during pollination and pollen tube growth are major mechanisms for pollen selection, the protein composition of the pollen surface or proteins secreted by the pollen may play a role. Those pollen surface proteins originate partly from sporophytic tissues from the anther, where the pollen is stored before release, such as oleopollenins which are known to be rapidly evolving (Schein et al. 2004). In a data set enriched for sporophytic pollen surface proteins (Yang et al. 2007), we observed higher levels of purifying selection and an increased rate of adaptive evolution ($f = 0.15$, $\alpha = 0.16$), even though the estimate for α was not significant, possibly do to the limited data availability.

As *A. thaliana* is highly self-fertilizing with low outcrossing rates, the genetic differences between male and female gametophytes of individual *A. thaliana* plants are therefore low and heterogeneity in selection may arise from two sources: First, *de novo* mutations originating from the microspore mother cells before the generative cell is formed. *De novo* mutations are thought to be negligible because of their rare occurrence. Second, from rare outcrossing events. In natural populations of *A. thaliana*, outcrossing rates can be as high as 15% (Bomblies et al. 2010), even though usual rates are around 1% or less. Therefore, even though outcrossing events are rare, their limited occurrence appears to be sufficient to generate a molecular pattern which is consistent with the consequences of sexual selection in *A. thaliana*.

The breakdown of self-incompatibility in *A. thaliana* evolved approximately 1 Ma (Tang et al. 2007), suggesting that polymorphisms mainly segregated under a regime of self-fertilization. As we used pairwise divergence estimates between *A. lyrata* and *A. thaliana*, d_n/d_s ratios mainly reflect the evolutionary history in a regime of outcrossing. Using additional sequence information from the *Capsella rubella* genome (Slotte et al. 2013), it is possible to obtain lineage-specific divergence estimates for a subset of genes for which orthologous sequences exist in each lineage. Using the *C. rubella* sequences, pairwise divergence estimates were not significantly different from lineage-specific estimates obtained for either *A. lyrata* or *A. thaliana* for the six reproductive tissues and random genes (table 1).

However, lineage-specific estimates for d_n/d_s in pollen and pollen tube genes are lower in the *A. lyrata*, but not in the *C. rubella* lineage when compared with *A. thaliana* (table 1). The lower median d_n/d_s values observed for *A. lyrata* may result either from increased purifying selection or reduced positive selection. In the first case, the DFE for *A. lyrata* is

Table 1. Median (in brackets 0.1 and 0.9 quantile) Estimates for d_n/d_s Using Ortholog Sequences from *Arabidopsis thaliana*, *A. lyrata*, and *Capsella rubella*.

Tissue	Median Pairwise d_n/d_s	Median Lineage-Specific d_n/d_s		
	A. thaliana versus *A. lyrata*	*A. thaliana*	*C. rubella*	*A. lyrata*
Female				
Egg cell	0.29 (0.05, 0.93)	0.28 (0.07, 1.12)	0.27 (0.06, 0.9)	0.26 (0.01, 3.7)
Synergids	0.27 (0.08, 0.94)	0.28 (0.08, 0.92)	0.28 (0.07, 0.8)	0.28 (0.06, 3.73)
Central cell	0.29 (0.07, 0.94)	0.3 (0.09, 1.12)	0.28 (0.07, 0.8)	0.25 (0.04, 2.01)
Male				
Pollen	0.14 (0.03, 0.46)	0.15 (0.02, 0.62)*	0.15 (0.02, 0.48)	0.12 (0.01, 0.58)*
Pollen tube	0.14 (0.03, 0.45)	0.15 (0.03, 0.61)*	0.15 (0.03, 0.47)	0.12 (0.01, 0.60)*
Sperm cell	0.2 (0.04, 0.62)	0.21 (0.04, 0.71)	0.2 (0.05, 0.54)	0.2 (0.03, 0.71)
Random	0.19 (0.03, 0.57)	0.18 (0.03, 0.7)	0.19 (0.04, 0.56)	0.2 (0.02, 0.77)

*$P < 0.001$ (*U* test, lineage-specific divergence *A. thaliana vs. A. lyrata*).

expected to be different from *A. thaliana* and *Capsella grandiflora*. As no whole-genome polymorphism data are currently available for *A. lyrata*, it is not possible to investigate this hypothesis further. Under the assumption that the DFE of pollen and pollen tube genes are similar in both *Arabidopsis* species, the level of adaptive evolution is reduced in *A. lyrata*. This is rather surprising, because pollen competition is supposed to be stronger in outcrossing species. Indeed, both increased purifying and diversifying selection act on pollen genes when compared with sperm genes in the self-incompatible species *C. grandiflora* (Arunkumar et al. 2013), similar to our observation in *A. thaliana*. However, Arunkumar et al. used expression data from *A. thaliana* and *Brassica napus* to define sets of sperm- and pollen-specific genes, and therefore focused on genes that are functionally conserved within Brassicaceae. *Capsella grandiflora* has undergone a substantial amount of adaptive evolution, which is possibly caused by positive selection facilitated by its high N_e (Slotte et al. 2010). There was little heterogeneity in d_n/d_s ratios between reproductive genes and random genes among the three lineages (table 1), which suggests that either the long-term effective population sizes are similar (Woolfit 2009) or that different effects of positive and negative selection lead to comparable d_n/d_s values. As the estimate of N_e based on nucleotide diversity is large in *C. grandiflora*, the effect of varying amounts of mutations that are effectively neutral are difficult to disentangle because the selective strength scales with N_e. In contrast, we not only investigated additional factors that may explain observed differences among gene sets such as recombination rate, expression level, and male–female differences, but also conclude that differential selection on sex-related genes explain patterns of variation.

Conclusion

The role of sexual selection in plants is highly debated, especially for hermaphrodites, which represent the most common type of reproduction among angiosperms (Moore and Pannell 2011). The majority of similar studies in animals have focused on differences in whole body or tissue-specific expression in male and female individuals and on contrasting patterns of gene sets located on sex chromosomes versus autosomes. Consequently, there is a great variety in definitions of sex specificity in animals. In *A. thaliana*, sex specificity is restricted to reproductive organs. We observed remarkable differences between genes expressed in pollen and pollen tubes to genes expressed in other reproductive tissues, with regard to their genomic location, gene age, sequence diversity, and interspecific divergence. Taken together, our findings suggest that selective forces acting on genes specific to pollen and pollen tubes are different in comparison with other reproductive tissues. This is likely the consequence of selective mechanisms acting on pollen and pollen tubes in the prezygotic stage of pollination by pollen competition and pollen tube–pistil interactions that suggest a plausible mechanism for sexual selection in *A. thaliana*.

Materials and Methods

Gene Expression Data Sets

Gene expression data were obtained from published sources (supplementary data S1, Supplementary Material online). These data comprise a variety of mixed and separate tissue types, and of specific cell types. Raw data were processed as described in Schmidt et al. (2011) except for an updated annotation of the ATH1 microarray (brainarray.mbni.med.umich.edu, TAIR version 14). In brief, data were RMA-normalized (Irizarry et al. 2003) and *P* values were calculated with *At*PANP. Genes with $P < 0.05$ were defined as being expressed (Wuest et al. 2010). Differential expression analysis was performed with limma (Smyth 2004) using an adjusted *P* value (false discovery rate) cutoff of 0.05 and a minimal fold-change of four (on a $\log_2$ scale). In total, we performed six tests for differential expression in which we compared sperm, egg, central cell, synergids, pollen, and pollen tube with the other tissues. To test for functional enrichment, we used goatools (https://github.com/tanghaibao/goatools, last accessed December 2, 2013). As a genome-wide reference, we randomly sampled 500 genes from the *A. thaliana* reference genome, which is comparable with the gene number of the three male tissues. We disregarded those genes that were included in the set of reproductive genes resulting in a set of 476 genes (random set). As a measurement of gene expression specificity, we obtained the τ index (Yanai et al. 2005) for each gene based on the available expression data. τ values range between 0 and 1 with larger values indicating higher specificity of the respective gene. As an approximation for expression intensity, we used the maximum normalized gene expression value of a particular gene in all tissues (Slotte et al. 2011).

Characteristics of Gene Sets and Identification of Segmentally Duplicated Blocks

To test whether reproductive genes show a significant deviation from a random chromosomal distribution, we compared the distribution of reproductive and randomly chosen genes over the five chromosomes and on a finer genomic scale of 10 Mb using χ^2 tests of independence. We also used information from the genome duplication database, based on synteny blocks within *A. thaliana* (Tang et al. 2008), to determine whether genes that are located within segmentally duplicated blocks and have a paralog in the corresponding segment originated from a WGD event in the history of the *A. thaliana* lineage, because those paralogs may show differences in their evolutionary rates (Gossmann and Schmid 2011).

Identification of Selective Effects

For each gene classified as reproductive gene, we obtained single-nucleotide polymorphism (SNP) data from 80 *A. thaliana* accessions (Cao et al. 2011) and, if available, estimated the divergence between the *A. thaliana* and *A. lyrata* (Hu et al. 2011) reference genomes. The pairwise divergence ($d_n/d_s = \omega$) was calculated with PAML, F3x4, runmode = −2. Lineage-specific estimates of d_n/d_s values

were calculated using additionally the *C. rubella* genome (Slotte et al. 2013) along with the free ratios branch model in PAML. We randomly sampled 20 accessions without replacement at each SNP position to speed up computing. To investigate whether the geographic population structure of the 80 accessions influenced the overall estimates, we conducted two subsequent analyses. First, to exclude the possibility that the sampling scheme inflated the number of polymorphisms segregating at low frequencies, we analyzed each of the eight geographic regions separately. Second, we used an alternative scheme by randomly sampling one allele from each of the eight populations at each SNP position (Wakeley 2001). It is possible to infer the proportion of mutations under purifying selection from the frequency distribution of mutations; for example, those mutations for which $N_e \times s > 1$, where N_e is the effective population size and s the mean selective effect (Keightley and Eyre-Walker 2007). We used the estimated frequency distribution of SNPs to infer the DFE of new mutations (Keightley and Eyre-Walker 2007) using DoFE 3.0 (http://www.lifesci.sussex.ac.uk/home/Adam_Eyre-Walker/Website/Software.html, last accessed December 2, 2013). For this analysis, we excluded singletons and summed data across genes for each tissue to speed up computing. The McDonald–Kreitman (MK) test (McDonald and Kreitman 1991) contrasts nonsynonymous to synonymous divergence (e.g., D_n/D_s) to diversity (e.g., P_n/P_s) to infer the proportion of adaptive substitutions α. We applied a derivative of the MK test which corrects for the impact of slightly deleterious mutations (Eyre-Walker and Keightley 2009). For each tissue analyzed, we contrasted polymorphism patterns of reproductive genes with available divergence estimates to obtain estimates of α and ω_a, which is the rate of adaptive substitution relative to the rate of synonymous substitution (Gossmann et al. 2010).

Gene Age and Recombination Rates

The rate of protein evolution of a gene may be determined by its age (Cai and Petrov 2010). To obtain a proxy for gene age, we used two different methods that are based on Blast to detect homologies. Simulations suggest that this approach works well for eukaryotes (Albá and Castresana 2007). For the first method, we used the most distant best pairwise Blast (Blastp, default parameters) hit from a set of ten species, which are in a phylogenetically hierarchical order (supplementary fig. S2, Supplementary Material online). If multiple genomes were available in a hierarchy level, we chose the one which we believed was annotated at the highest quality. For the second method, we used gene age estimates by Quint et al. (2012), which are based on the construction of a phylostratigraphic map. In principle, this method uses one-way Blast hits against a set of >1,000 genomes, which can be assigned to 1 of 12 phylostrata. For consistency, we define for both methods that youngest genes have the gene age 0 (specific to *A. thaliana*) and older genes have a positive value. For recombination rate variation across the *A. thaliana* genome, we obtained population recombination rate ($\rho = 4N_e r$, where r is the recombination rate per generation) estimates from 1,307 worldwide accessions (Horton et al. 2012).

Supplementary Material

Supplementary data S1, figures S1–S7, and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank two anonymous referees for their comments which have helped to improve the manuscript. They thank Ramesh Arunkumar, Stephen Wright, and colleagues for providing an early draft of their manuscript on tissue-specific patterns of selection in *C. grandiflora*. They also thank Kai Zeng and Jessica Stapley for comments on an earlier version of the manuscript. This work was supported by the Deutsche Forschungsgemeinschaft (EVOREP project SCHM1354/7-1) to K.J.S., the University of Zürich, and a grant of the Swiss National Science Foundation to U.G.

References

Albá MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. *BMC Evol Biol.* 7:53.

Arunkumar R, Josephs EB, Williamson RJ, Wright SI. 2013. Pollen-specific, but not sperm-specific genes show stronger purifying selection and higher rates of positive selection than sporophytic genes in *Capsella grandiflora*. *Mol Biol Evol.* 30:2475–2486.

Baines JF, Sawyer SA, Hartl DL, Parsch J. 2008. Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in *Drosophila*. *Mol Biol Evol.* 25:1639–1650.

Bernasconi G, Ashman TL, Birkhead TR, Bishop JD, Grossniklaus U, Kubli E, Marshall DL, Schmid B, Skogsmyr I, Snook RR, et al. 2004. Evolutionary ecology of the prezygotic stage. *Science* 303:971–975.

Bomblies K, Yant L, Laitinen RA, Kim ST, Hollister JD, Warthmann N, Fitz J, Weigel D. 2010. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genet.* 6:e1000890.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in *Arabidopsis*. *Nature* 416:531–534.

Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol Evol.* 2: 393–409.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet.* 43: 956–963.

Carlson AL, Gerald JNF, Telligman M, Roshanmanesh J, Swanson RJ. 2011. Defining the genetic architecture underlying female- and male-mediated nonrandom mating and seed yield traits in *Arabidopsis*. *Plant Physiol.* 157:1956–1964.

Carlson AL, Telligman M, Swanson RJ. 2009. Incidence and post-pollination mechanisms of nonrandom mating in *Arabidopsis thaliana*. *Sex Plant Reprod.* 22:257–262.

Chapman LA, Goring DR. 2010. Pollen-pistil interactions regulating successful fertilization in the Brassicaceae. *J Exp Bot.* 61:1987–1999.

Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am Naturalist.* 130:113–146.

Charlesworth D. 2002. Plant sex determination and sex chromosomes. *Heredity* 88:94–101.

Clark NL, Swanson WJ. 2005. Pervasive adaptive evolution in primate seminal proteins. *PLoS Genet.* 1:e35.

Cutter AD, Ward S. 2005. Sexual and temporal dynamics of molecular evolution in *C. elegans* development. *Mol Biol Evol.* 22:178–188.

Dellaporta SL, Calderon-Urrea A. 1993. Sex determination in flowering plants. *Plant Cell* 5:1241–1251.
Dickinson H, Grant-Downton R. 2009. Bridging the generation gap: flowering plant gametophytes and animal germlines reveal unexpected similarities. *Biol Rev Camb Philos Soc.* 84:589–615.
Dorus S, Wasbrough ER, Busby J, Wilkin EC, Karr TL. 2010. Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes. *Mol Biol Evol.* 27:1235–1246.
Dresselhaus T, Sprunck S. 2012. Plant fertilization: maximizing reproductive success. *Curr Biol.* 22:R487–R489.
Ellegren H. 2007. Characteristics, causes and evolutionary consequences of male-biased mutation. *Proc Biol Sci.* 274:1–10.
Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet.* 8:689–698.
Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol.* 26: 2097–2108.
Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227–1234.
Filatov DA, Charlesworth D. 2002. Substitution rates in the X- and Y-linked genes of the plants, *Silene latifolia* and *S. dioica*. *Mol Biol Evol.* 19:898–907.
Foxe JP, Dar VU, Zheng H, Nordborg M, Gaut BS, Wright SI. 2008. Selection on amino acid substitutions in *Arabidopsis*. *Mol Biol Evol.* 25:1375–1383.
Gnad F, Parsch J. 2006. Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. *Bioinformatics* 22:2577–2579.
Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol Evol.* 4:658–667.
Gossmann TI, Schmid KJ. 2011. Selection-driven divergence after gene duplication in *Arabidopsis thaliana*. *J Mol Evol.* 73:153–165.
Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome-wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol.* 27:1822–1832.
Gossmann TI, Woolfit M, Eyre-Walker A. 2011. Quantifying the variation in the effective population size within a genome. *Genetics* 189: 1389–1402.
Grath S, Parsch J. 2012. Rate of amino acid substitution is influenced by the degree and conservation of male-biased transcription over 50 myr of *Drosophila* evolution. *Genome Biol Evol.* 4:346–359.
Grossniklaus U. 2011. Plant germline development: a tale of cross-talk, signaling, and cellular interactions. *Sex Plant Reprod.* 24:91–95.
Gschwend AR, Yu Q, Tong EJ, Zeng F, Han J, VanBuren R, Aryal R, Charlesworth D, Moore PH, Paterson AH, et al. 2012. Rapid divergence and expansion of the X chromosome in papaya. *Proc Natl Acad Sci U S A.* 109:13716–13721.
Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res.* 8:269–294.
Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, Muliyati NW, Platt A, Sperone FG, Vilhjálmsson BJ, et al. 2012. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat Genet.* 44:212–216.
Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet.* 43:476–481.
Hülskamp M, Schneitz K, Pruitt RE. 1995. Genetic evidence for a long-range activity that directs pollen tube guidance in *Arabidopsis*. *Plant Cell* 7:57–64.
Hurst LD, Ellegren H. 1998. Sex biases in the mutation rate. *Trends Genet.* 14:446–452.
Irizarry R, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, Speed T. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249.
Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* 13:R3.
Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177: 2251–2261.
Kessler SA, Grossniklaus U. 2011. She's the boss: signaling in pollen tube reception. *Curr Opin Plant Biol.* 14:622–627.
Liu SL, Baute GJ, Adams KL. 2011. Organ and cell type-specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in *Arabidopsis thaliana*. *Genome Biol Evol.* 3:1419–1436.
Ma H, Sundaresan V. 2010. Development of flowering plant gametophytes. *Curr Top Dev Biol.* 91:379–412.
Mank JE, Hultin-Rosenberg L, Axelsson E, Ellegren H. 2007. Rapid evolution of female-biased, but not male-biased, genes expressed in the avian brain. *Mol Biol Evol.* 24:2698–2706.
Mank JE, Vicoso B, Berlin S, Charlesworth B. 2010. Effective population size and the Faster-X effect: empirical results and their interpretation. *Evolution* 64:663–674.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652–654.
Meisel RP. 2011. Towards a more nuanced understanding of the relationship between sex-biased gene expression and rates of protein-coding sequence evolution. *Mol Biol Evol.* 28: 1893–1900.
Ming R, Bendahmane A, Renner SS. 2011. Sex chromosomes in land plants. *Annu Rev Plant Biol.* 62:485–514.
Müller L, Grath S, von Heckel K, Parsch J. 2012. Inter- and intraspecific variation in *Drosophila* genes with sex-biased expression. *Int J Evol Biol.* 2012:963976.
Moore JC, Pannell JR. 2011. Sexual selection in plants. *Curr Biol.* 21: R176–R182.
Paape T, Bataillon T, Zhou P, Kono TJY, Briskine R, Young ND, Tiffin P. 2013. Selection, genome-wide fitness effects and evolutionary rates in the model legume *Medicago truncatula*. *Mol Ecol.* 22: 3525–3538.
Palanivelu R, Tsukamoto T. 2011. Pathfinding in angiosperm reproduction: pollen tube guidance by pistils ensures successful double fertilization. *Wiley Interdiscip Rev Dev Biol.* 1:96–113.
Parsch J, Ellegren H. 2013. The evolutionary causes and consequences of sex-biased gene expression. *Nat Rev Genet.* 14:83–87.
Price CS, Dyer KA, Coyne JA. 1999. Sperm competition between *Drosophila* males involves both displacement and incapacitation. *Nature* 400:449–452.
Pröschel M, Zhang Z, Parsch J. 2006. Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. *Genetics* 174: 893–900.
Quint M, Drost HG, Gabel A, Ullrich KK, Bönn M, Grosse I. 2012. A transcriptomic hourglass in plant embryogenesis. *Nature* 490: 98–101.
Rice WR. 1996. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature* 381:232–234.
Schein M, Yang Z, Mitchell-Olds T, Schmid KJ. 2004. Rapid evolution of a pollen-specific oleosin-like gene family from *Arabidopsis thaliana* and closely related species. *Mol Biol Evol.* 21:659–669.
Schmid MW, Schmidt A, Klostermeier UC, Barann M, Rosenstiel P, Grossniklaus U. 2012. A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing. *PLoS One* 7:e29685.
Schmidt A, Schmid MW, Grossniklaus U. 2012. Analysis of plant germline development by high-throughput RNA profiling: technical advances and new insights. *Plant J.* 70:18–29.
Schmidt A, Wuest SE, Vijverberg K, Baroux C, Kleen D, Grossniklaus U. 2011. Transcriptome analysis of the *Arabidopsis* megaspore mother cell uncovers the importance of RNA helicases for plant germline development. *PLoS Biol.* 9:e1001155.

Slotte T, Bataillon T, Hansen TT, Onge KS, Wright SI, Schierup MH. 2011. Genomic determinants of protein evolution and polymorphism in *Arabidopsis*. *Genome Biol Evol.* 3:1210–1219.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Mol Biol Evol.* 27: 1813–1821.

Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, Steige K, Platts AE, Escobar JS, Newman LK, et al. 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet.* 45:831–835.

Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415:1022–1024.

Smyth GK. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* 3:Article3.

Spillane C, Schmid KJ, Laoueillé-Duprat S, Pien S, Escobar-Restrepo JM, Baroux C, Gagliardini V, Page DR, Wolfe KH, Grossniklaus U. 2007. Positive Darwinian selection at the imprinted *MEDEA* locus in plants. *Nature* 448:349–352.

Takeuchi H, Higashiyama T. 2011. Attraction of tip-growing pollen tubes by the female gametophyte. *Curr Opin Plant Biol.* 14: 614–621.

Tang C, Toomajian C, Sherman-Broyles S, Plagnol V, Guo YL, Hu TT, Clark RM, Nasrallah JB, Weigel D, Nordborg M. 2007. The evolution of selfing in *Arabidopsis thaliana*. *Science* 317: 1070–1072.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. *Science* 320:486–488.

Twell D. 2011. Male gametogenesis and germline specification in flowering plants. *Sex Plant Reprod.* 24:149–160.

Vicoso B, Charlesworth B. 2009. Effective population size and the faster-X effect: an extended model. *Evolution* 63:2413–2426.

Wakeley J. 2001. The coalescent in an island model of population subdivision with variation among demes. *Theor Popul Biol.* 59: 133–144.

Whittle CA, Johnston MO. 2002. Male-driven evolution of mitochondrial and chloroplastidial DNA sequences in plants. *Mol Biol Evol.* 19: 938–949.

Whittle CA, Johnston MO. 2003. Male-biased transmission of deleterious mutations to the progeny in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A.* 100:4055–4059.

Williams JH. 2008. Novelties of the flowering plant pollen tube underlie diversification of a key life history stage. *Proc Natl Acad Sci U S A.* 105:11259–11263.

Woolfit M. 2009. Effective population size and the rate and pattern of nucleotide substitutions. *Biol Lett.* 5:417–420.

Wuest SE, Vijverberg K, Schmidt A, Weiss M, Gheyselinck J, Lohr M, Wellmer F, Rahnenführer J, von Mering C, Grossniklaus U. 2010. *Arabidopsis* female gametophyte gene expression map reveals similarities between plant and animal gametes. *Curr Biol.* 20: 506–512.

Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304–309.

Yampolsky C, Yampolsky H. 1922. Distribution of sex forms in the phanerogamic flora. *Bibl Genet.* 3:1–62.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21: 650–659.

Yang C, Vizcay-Barrena G, Conner K, Wilson ZA. 2007. *MALE STERILITY1* is required for tapetal development and pollen wall biosynthesis. *Plant Cell* 19:3530–3548.

Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Mol Biol Evol.* 28:2359–2369.

8.5 Analysis of plant germline development by high-throughput RNA profiling: technical advances and new insights

The following review is published in "The Plant Journal" (published by Blackwell Publishing Ltd, all rights reserved)[1]. I created all figures and, where applicable, performed the underlying data analysis. I further provided a draft for the section "RNA-Seq outperforms microarrays in terms of detection range and for transcriptome profiling of non-model species".

[1]Schmidt, A, Schmid, MW, and Grossniklaus, U (2012) Analysis of plant germline development by high-throughput RNA profiling: technical advances and new insights. The Plant Journal 70: 18–29.

the plant journal

SEB Society for Experimental Biology

The Plant Journal (2012) **70**, 18–29 doi: 10.1111/j.1365-313X.2012.04897.x

HIGH-RESOLUTION MEASUREMENTS IN PLANT BIOLOGY

Analysis of plant germline development by high-throughput RNA profiling: technical advances and new insights

Anja Schmidt*, Marc W. Schmid and Ueli Grossniklaus

Institute of Plant Biology and Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, 8008 Zürich, Switzerland

Received 17 November 2011; revised 20 December 2011; accepted 22 December 2011.
*For correspondence (fax + 41 44 6348204; e-mail aschmidt@botinst.uzh.ch).

SUMMARY

Reproduction is a crucial step in the life cycle of plants. The male and female germline lineages develop in the reproductive organs of the flower, which in higher plants are the anthers and ovules, respectively. Development of the germline lineage initiates from a dedicated sporophytic cell that undergoes meiosis to form spores that subsequently give rise to the gametophytes through mitotic cell divisions. The mature male and female gametophytes harbour the male (sperm cells) and female gametes (egg and central cell), respectively. Those unite during double fertilization to initiate embryo and endosperm development in sexually reproducing higher plants. While cytological changes involved in development of the germline lineages have been well characterized in a number of species, investigation of the transcriptional basis underlying their development and the specification of the gametes proved challenging. This is largely due to the inaccessibility of the cells constituting the germline lineages, which are enclosed by sporophytic tissues. Only recently, these technical limitations could be overcome by combining new methods to isolate the relevant cells with powerful transcriptional profiling methods, such as microarrays or high-throughput sequencing of RNA. This review focuses on these technical advances and the new insights gained from them concerning the transcriptional basis and molecular mechanisms underlying germline development.

Keywords: germline development, cell-type-specific isolation, microarrays, laser-assisted microdissection (LAM), fluorescence-activated cell sorting (FACS), transcriptomics, RNA-Seq.

THE PLANT LIFE CYCLE AND GERMLINE DEVELOPMENT

Unlike in animals, the precursors of the plant germline lineage are not set aside early during development (reviewed by Dickinson and Grant-Downton, 2009). In contrast, the plant life cycle alternates between a diploid sporophytic and a haploid gametophytic generation. The gametophytic generation has been progressively reduced during the evolution of land plants (reviewed by Haig and Wilczek, 2006; Dickinson and Grant-Downton, 2009). In bryophytes, the gametophytic generation is the dominant multicellular phase, while the life cycle of pteridophytes and higher land plants is dominated by the sporophyte and the gametophyte remains nutritionally dependent on the sporophyte in the latter (Haig and Wilczek, 2006). In flowering plants (angiosperms), gametophyte development takes place in specialized reproductive organs of the flower, the anthers and the ovules, respectively. Single determined sporophytic cells, the archespores, get selected to undergo meiosis and to give rise to dimorphic male and female spores during micro- and megasporogenesis, respectively. Subsequently, the haploid gametophytes harbouring the gametes develop from the spores through mitotic divisions. Thus, the archesporial cells can be viewed as the first cells of the plant reproductive or germline lineages committed to produce the gametes (Grossniklaus, 2011; Schmidt *et al.*, 2011). However, later stages of reproductive development, e.g. when the gamete lineage is specified, have also been proposed to be the decisive step in germline specification (Dickinson and Grant-Downton, 2009; Twell, 2011).

While the development of both male and female reproductive lineages starts with the selection of a sporophytic cell (meiocyte) that undergoes meiosis to form haploid spores, there are important differences during gametophyte

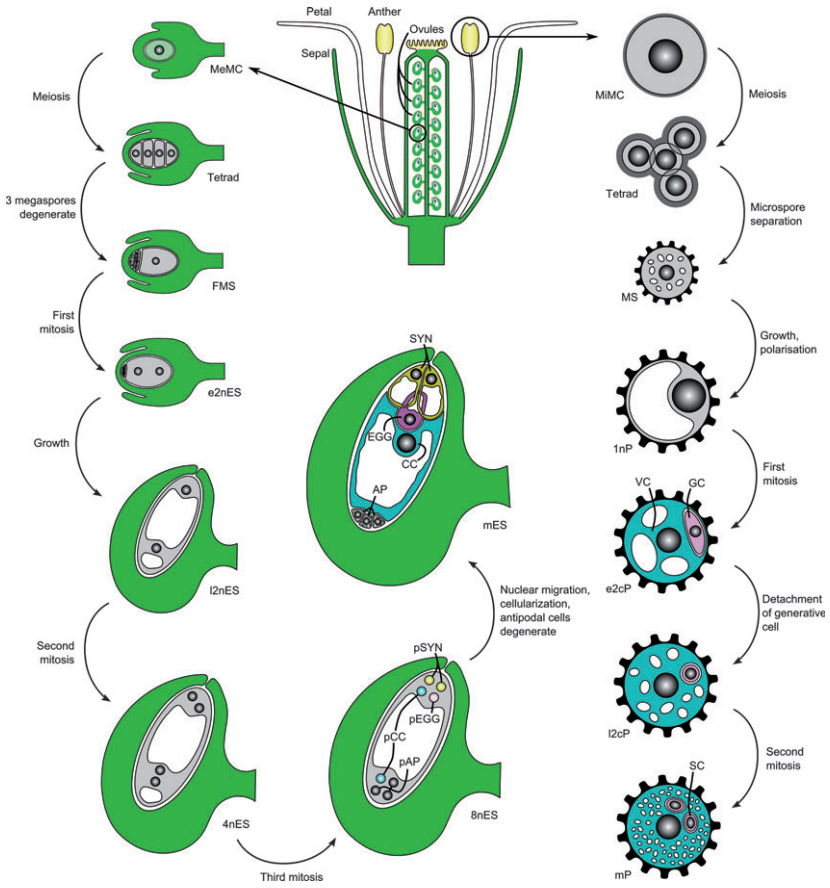


Figure 1. Schematic representation of male and female gametophyte development in *Arabidopsis thaliana*.
Germline development starts with sporophytic cells differentiating in to spore mother cells (female, megaspore mother cell; male, microspore/pollen mother cell). These mother cells undergo meiosis resulting in the formation of four haploid spores, three of which degenerate in the case of the female. Each functional spore subsequently gives rise to one gametophyte. Abbreviations: MeMC, megaspore mother cell; FMS, functional megaspore (FG1); e2nES/l2nES, early/late two-nucleate embryo sac (FG2/FG3); 4nES, four-nucleate embryo sac (FG4); 8nES, eight-nucleate embryo sac (FG5–FG6); (p)AP, (precursor of) antipodal cell; (p)EGG, (precursor of) egg cell; (p)SYN, (precursor of) synergid cell; (p)CC, (precursor of) central cell; mES, mature embryo sac (FG7); MiMC, microspore mother cell; MS, microspore; 1nP, uninucleate pollen; e2cP/l2cP, early/late bicellular pollen; mP, mature pollen; VC, vegetative cell; GC, generative cell; SC, sperm cell.

development and gamete specification (Figure 1). In the male germline, all four microspores originating from a pollen or microspore mother cell (MiMC) undergo asymmetric pollen mitosis I (PMI) to produce a vegetative and a generative cell (reviewed by Borg *et al.*, 2009). During pollen mitosis II (PMII) the generative cell divides to form two sperm cells that are delivered to the female gametophyte by the pollen tube, which is formed by growth of the vegetative cell (Figure 1). The timing of these developmental processes varies between species. In most species, the generative cell undergoes PMII during pollen tube growth. In other species, like crucifers or grasses, however, PMII and maturation of the tricellular pollen takes place in the anther prior to pollen release and germination (Boavida *et al.*, 2005). During development of the female germline, a *Polygonum*-type embryo sac is formed in >70% of all species, including Brassicaceae (e.g. the model plant *Arabidopsis thaliana*) and Gramineae (e.g. maize, wheat, rice) (reviewed by Yadegari and Drews, 2004; Brukhin *et al.*, 2005; Sprunck and Gross-Hardt, 2011). In contrast to the male germline, however, typically only one megaspore survives after meiosis of the megaspore mother cell (MeMC) while the others degenerate. This functional megaspore gives rise to a *Polygonum* female gametophyte (embryo sac) by three rounds of mitosis in a syncytium, followed by cellularization of the eight-nucleate embryo sac (Figure 1). The embryo sac harbours the female gametes, the egg and central cell, both of which require fertilization to initiate development of the diploid embryo and the triploid endosperm, respectively (Figure 1).

The identification of genes important for plant reproduction, and especially the investigation of the gene regulatory networks underlying plant germline specification and development, have proved to be difficult due to technical obstacles. For the female reproductive lineage in particular, the low number of cells in the germline lineage and their inaccessibility – they develop enclosed by sporophytic tissue – have long hampered transcriptional profiling approaches. However, recent advances in establishing methods to isolate individual cells from the germline lineages in combination with high-throughput transcriptome profiling techniques have yielded important new insights and will be summarized in this review.

EARLY APPROACHES TO IDENTIFY GENES IMPORTANT FOR OR EXPRESSED DURING PLANT GERMLINE DEVELOPMENT

Over the last decade, genetic screens using the model plant *A. thaliana* have led to the identification of a number of genes involved in sporo- and gametogenesis (reviewed in Yadegari and Drews, 2004; Brukhin *et al.*, 2005; Berger and

Twell, 2011; Chang *et al.*, 2011; Sprunck and Gross-Hardt, 2011; Twell, 2011). But only a few genes with specific expression in gametophytic cells were identified, mostly by enhancer detection, a method allowing the identification of genes based on their pattern of expression (Sundaresan *et al.*, 1995; Grossniklaus *et al.*, 1998, 2002; Gross-Hardt *et al.*, 2007). A more comprehensive picture of the transcriptional landscape in the cells of the gametophyte only became possible with the advent of transcript profiling methods. These were used in combination with mutants lacking a female gametophyte, such as *sporocyteless*/*nozzle* (*spl*/*nzz*), *coatlique* (*coa*) and *determinant infertile 1* (*dif1*), in which development typically arrests before the initiation of meiosis, at the megaspore stage, and during meiosis, respectively (Bai *et al.*, 1999; Bhatt *et al.*, 1999; Yang *et al.*, 1999; Johnston *et al.*, 2007). Transcriptional profiling of isolated ovules or pistils from those mutants was subsequently used to identify genes expressed in the female gametophyte. Comparative profiling of wild-type and *spl* mutant ovules using Affymetrix ATH1 arrays identified 225 potentially gametophyte expressed genes (Yu *et al.*, 2005), while 1260 potentially embryo sac expressed genes were identified based on comparative profiling of wild-type and mutant *spl* and *coa* ovules and pistils, respectively (Johnston *et al.*, 2007). In addition, 71 and 382 genes were identified to be downregulated in *dif1* mutant ovules (Jones-Rhoades *et al.*, 2007; Steffen *et al.*, 2007), using either the Affymetrix ATH1 or tiling arrays. While giving important new insights into the transcriptional basis of embryo sac development, these studies were limited by several drawbacks: (i) expression of a considerable number of gametophytic genes was superposed by the expression of sporophytic genes in pistils or ovules and thus remained undetected, (ii) different influences of the mutants on sporophytic gene expression made the results more difficult to interpret, and (iii) the numbers of genes potentially expressed in the embryo sac remained considerably smaller than the estimated transcriptome size required for germline development (Yu *et al.*, 2005; Johnston *et al.*, 2007; Jones-Rhoades *et al.*, 2007; Steffen *et al.*, 2007). It became obvious from these studies that significant technical improvements were required to allow germline-specific sampling and transcriptional profiling (Johnston *et al.*, 2007).

TECHNICAL ADVANCES IN ISOLATION OF GERMLINE-SPECIFIC CELLS AND TISSUES

The first attempts to isolate embryo sacs from ovule tissues date back to the middle of the 20th century (reviewed by Xin and Sun, 2010). Isolation of male and female gametes in combination with expressed sequence tag (EST) sequencing and generation of cDNA libraries subsequently allowed the identification of genes expressed in gametes from maize (*Zea mays*), wheat (*Triticum aestivum*), *Nicotiana tabacum* and Arabidopsis, and from the generative cell from *Lilium longiflorum* (Dresselhaus *et al.*, 1994; Kumlehn *et al.*, 2001; Xu *et al.*, 2002; Engel *et al.*, 2003; Lê *et al.*, 2005; Sprunck *et al.*, 2005; Okada *et al.*, 2006; Yang *et al.*, 2006; Xin and Sun, 2010; Xin *et al.*, 2011). However, suitable techniques for targeted isolation of almost every cell type of interest from male and female germline lineages were established only recently, based on micromanipulation, fluorescence-activated cell sorting (FACS), and laser-assisted microdissection (LAM) (reviewed by Xin and Sun, 2010; Hu *et al.*, 2011). In addition, a method for the isolation of nuclei from specific cell types (isolation of nuclei tagged in specific cell types, INTACT) has recently been developed (Deal and Henikoff, 2011). In brief, micromanipulation is based on manual dissection of the tissue, sometimes in combination with enzymatic digestions of cell wall components, while FACS sorts cells based on their fluorescence and light scattering characteristics (reviewed by Hu *et al.*, 2011). During LAM, cells or tissue types of interest are isolated with a laser from thin sections of fixed and embedded tissue (reviewed in Day *et al.*, 2005; Nelson *et al.*, 2006). The method was originally developed for the isolation of specific cells from animal tissues (Emmert-Buck *et al.*, 1996) and first used for plant cells only more recently (Kerk *et al.*, 2003; Casson *et al.*, 2005). The INTACT method, on the other hand, uses affinity-based purification of nuclei expressing biotinylated proteins in the nuclear envelope of the target cells (Deal and Henikoff, 2011). The suitability of the different methods for the isolation of individual cells from male and female germline lineages, however, is largely dependent on the cell type of interest and the species used. Male gametophytic cells can be relatively easily isolated, e.g. by osmotic shock and separation by Percoll gradient centrifugation, as successfully applied for uninucleate microspores, binucleate pollen, and sperm cells from Arabidopsis and rice (*Oryza sativa*) (Table 1; Honys and Twell, 2004; Wei *et al.*, 2010; reviewed by Xin and Sun, 2010). In addition, FACS has been successfully used to sort mature Arabidopsis pollen and to isolate sperm cells from maize and Arabidopsis (Table 1). Micromanipulation has been applied to isolate male Arabidopsis meiocytes (microspore mother cells, MiMCs) (Table 1) and the generative cell of *L. longiflorum*, but also embryo sacs, female gametes and zygotes from a variety of different species including maize, *A. thaliana*, *O. sativa*, *Tourenia fournieri* and *Alstroemeria aurea* (Becker *et al.*, 2003; Engel *et al.*, 2003; Pina *et al.*, 2005; Hoshino *et al.*, 2006; Okada *et al.*, 2006, 2007; Chen *et al.*, 2010; Takanashi *et al.*, 2010; Ohnishi *et al.*, 2011; Yang *et al.*, 2011; Libeau *et al.*, 2011; reviewed by Xin and Sun, 2010; Hu *et al.*, 2011). Disadvantages of these powerful techniques, however, are that FACS often requires the use of a cell-type- or tissue-specific marker, as does the INTACT method. Apart from this, FACS, INTACT and micromanipulation may require prolonged handling or treatment with macerating enzymes, such that effects on RNA expression patterns or RNA

Table 1 Summary of recent transcriptome analyses of different developmental stages from pollen mother cell and megaspore mother cell to mature gametophytes of the male and female plant germline lineages using microarrays or high-throughput sequencing of RNA (RNA-Seq). Studies analysing exclusively stages after pollen germination or fertilization are not included

Developmental stage	Isolation technique	Species	Transcriptome profiling method	Literature
Male germline lineage				
Pre-meiotic pollen mother cell	Laser microdissection	*Oryza sativa* ssp. *japonica*	44K Agilent microarray	Tang *et al.* (2010)
Meiocyte	Micromanipulation	*Arabidopsis thaliana*	Solid sequencing	Yang *et al.* (2011)
Meiocyte	Micromanipulation	*A. thaliana*	Illumina sequencing	Chen *et al.* (2010)
Meiocyte	Micromanipulation	*A. thaliana*	CATMA microarray	Libeau *et al.* (2011)
Meiocyte, tetrad, UNM, BCP, TCP	Laser microdissection	*O. sativa* ssp. *japonica*	44K Agilent microarray	Suwabe *et al.* (2008), Hobo *et al.* (2008), Hirano *et al.* (2008)
UNM, BCP, TCP	Percoll gradient centrifugation	*A. thaliana*	Affymetrix ATH1 array	Honys and Twell (2004)
UNM, BCP, TCP	Percoll gradient centrifugation	*O. sativa* ssp. *japonica*	Affymetrix rice genome array	Wei *et al.* (2010)
Generative cell	Micromanipulation	*Lilium longiflorum*	cDNA microarrays	Okada *et al.* (2007)
Mature pollen		*A. thaliana*	Affymetrix ATH1 array	Schmid *et al.* (2005)
Mature pollen (hydrated, non hydrated)	FACS	*A. thaliana*	8K Affymetrix GeneCHIP	Becker *et al.* (2003)
Mature pollen	Filtration	*A. thaliana*	8K Affymetrix GeneCHIP	Honys and Twell (2003)
Mature pollen	FACS	*A. thaliana*	Affymetrix ATH1 array	Pina *et al.* (2005)
Mature pollen	Manual collection after rubbing anthers together on cover slips	*Glycine max*	Soybean GeneCHIP	Haerizadeh *et al.* (2009)
Mature pollen, anthers	Collected from anthers shedding pollen	*Z. mays*	44K maize oligonucleotide array	Ma *et al.* (2008)
Pollen germination, pollen tube growth	Vacuum method	*A. thaliana*	Affymetrix ATH1 array	Wang *et al.* (2008)
Pollen, pollen tubes	Separation from anthers with steel sieve	*Petunia axillaris*	cDNA spotted microarray	Ishimizu *et al.* (2010)
Pollen, pollen tubes (germinated *in vitro* and *semi in vivo*)	Vacuum method	*A. thaliana*	ATH1 array	Qin *et al.* (2009)
Sperm cells	FACS	*A. thaliana*	Affymetrix ATH1 array	Borges *et al.* (2008)
Sperm cells	Morphology based selection	*Plumbago zeylanica*	cDNA spotted microarray	Gou *et al.* (2009)
Female germline lineage				
Megaspore mother cell (MeMC)	Laser microdissection	*A. thaliana*	Affymetrix ATH1 array	Schmidt *et al.* (2011)
Egg cell, central cell, synergids	Laser microdissection	*A. thaliana*	Affymetrix ATH1 array	Wuest *et al.* (2010)
Central cell	Laser microdissection	*A. thaliana*	RNA-seq, SOLiD	Schmid *et al.* (2012)
Egg cell, synergid cells	Micromanipulation	*O. sativa* ssp. *japonica*	44K Agilent microarray	Ohnishi *et al.* (2011)

UNM, uninucleate microspore; BCP, bicellular pollen; TCP, tricellular pollen; FACS, fluorescence-activated cell sorting.

stability cannot be fully excluded. However, if handling time is kept short, the transcritional program of specific cell types does not appear to undergo substantial changes (Birnbaum *et al.*, 2003). LAM, on the other hand, is applicable for cell-type-specific isolation with little cross-contamination for a variety of purposes, because: (i) the use of specific markers is not required, and (ii) the tissue is fixed prior to any manipulation, such that transcriptional profiles are unaffected by the handling (Wuest *et al.*, 2010; Schmidt *et al.*, 2011; Schmid *et al.*, 2012). The downside of LAM is that it can be very time-consuming, depending on the cell type of interest. Also, the cell type of interest needs to be structurally distinguishable from the surrounding tissue in thin sections and isolated cells may contain minor contamination from neighbouring cells depending on the exact structural organization of the tissue within the section. While the laser beam leaves nucleic acids in the adjacent cytoplasm mostly intact, the thickness of the beam/section, and thus the suitability for the isolation of small cell types, varies with the LAM system used. Laser-assisted microdissection has been successfully applied to profile the cell-type-specific transcriptomes at different developmental stages of the male and female germline in different species, including MeMCs (Schmidt *et al.*, 2011), the three cell types of mature female gametophytes in Arabidopsis (Wuest *et al.*, 2010) and different developmental stages of the male germline in rice (*O. sativa* ssp. *japonica* 'Nipponbare'), including pre-meiotic MiMCs, microspores, bicellular and tricellular pollen (Hirano *et al.*, 2008; Hobo *et al.*, 2008; Suwabe *et al.*, 2008; Tang *et al.*, 2010).

INCREASING ESTIMATES OF TRANSCRIPTOME SIZE OF THE GERMLINE LINEAGES REFLECT ADVANCES IN PROFILING METHODS

The relative ease of access to and isolation of cells from the male as compared to the female germline and their higher abundance is reflected by the considerably higher number of studies analysing gene expression of the male as compared to the female germline (Table 1, Figures 2 and 3). This is largely due to the number of cells that can be isolated in a certain time period and, consequently, the amount of total RNA that can be obtained for transcriptional studies. For example, around 100 000 Arabidopsis sperm cells could be isolated in one FACS session, and 16 ng of total RNA was used as input for subsequent transcriptome analyses (Borges *et al.*, 2008). In addition, approximately 480 MiMCs per Arabidopsis flower could be isolated and 3.5 μg of total RNA was obtained from samples of approximately 57 600 cells (Libeau *et al.*, 2011). In contrast, one Arabidopsis flower harbours only about 50 ovules with one developing female germline lineage each. Using micromanipulation of target cells from rice, 3000 egg cells and 1000 synergid cells were collected (Ohnishi *et al.*, 2011). In Arabidopsis, several hundred cells from the female germline can be isolated

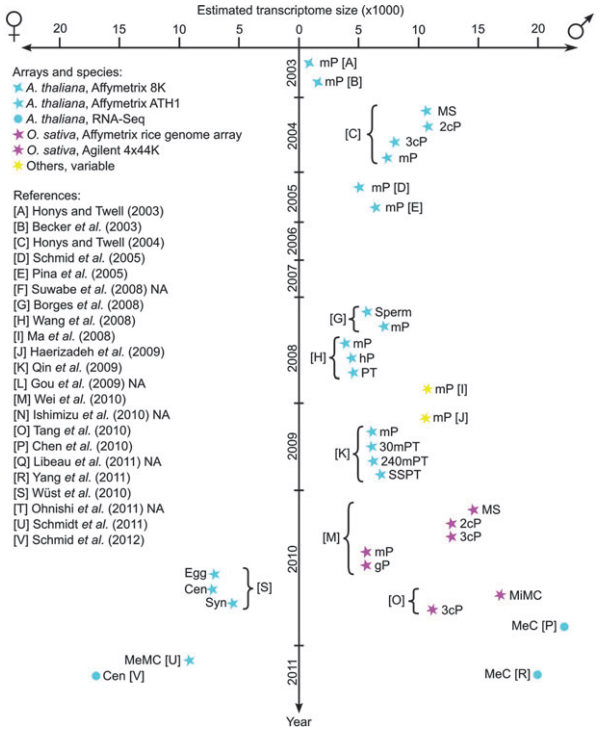


Figure 2. Advances in transcriptional profiling of gametophytes and gametophytic cell types.
The figure summarizes the transcriptome sizes of various stages of different cell and tissue types of the germline from several angiosperms estimated by high-throughput profiling (arrays or high-throughput RNA sequencing). The example of *Arabidopsis thaliana* illustrates the technical advances in transcriptional profiling with a general increase of transcriptome size estimates in the past few years. Also visible is the reduction of transcriptome size during pollen development. Abbreviations: MeMC, megaspore mother cell; egg, egg cell; syn, synergid cell; cen, central cell; MiMC, microspore mother cell; MeC, meiocyte; MS, microspore; 2cP, bicellular pollen; 3cP, tricellular pollen; mP, mature pollen; hP, hydrated pollen; gP, germinated pollen; PT, pollen tube; 30mPT/240mPT, pollen tube grown for 30/240 min; SSPT, pollen tube grown through stigma and style. References marked with NA do not provide transcriptome sizes.

separately using LAM, resulting in an estimated 0.3–1.5 ng of isolated total RNA (Wuest *et al.*, 2010; Schmidt *et al.*, 2011; Schmid *et al.*, 2012). Due to the small amounts of total RNA yielded from the isolation of specfic cells of the female germline, linear amplification of the mRNA is required prior to transcriptome analysis (Wuest *et al.*, 2010; Schmidt *et al.*, 2011; Schmid *et al.*, 2012), typically resulting in a shortening of the RNA fragments and a preferential amplification of, on average, approximately 400–500 bp of 3′ sequences of the transcripts. To account for the cell-type-specific analysis in combination with this amplification bias, the *At*PANP algorithm has been developed and tested for the analysis of Affymetrix ATH1 array data, outperforming the standard MAS5 algorithm in terms of accuracy and precision (Wuest *et al.*, 2010).

First studies to investigate transcriptomes of plant gametophytes using microarrays were performed using

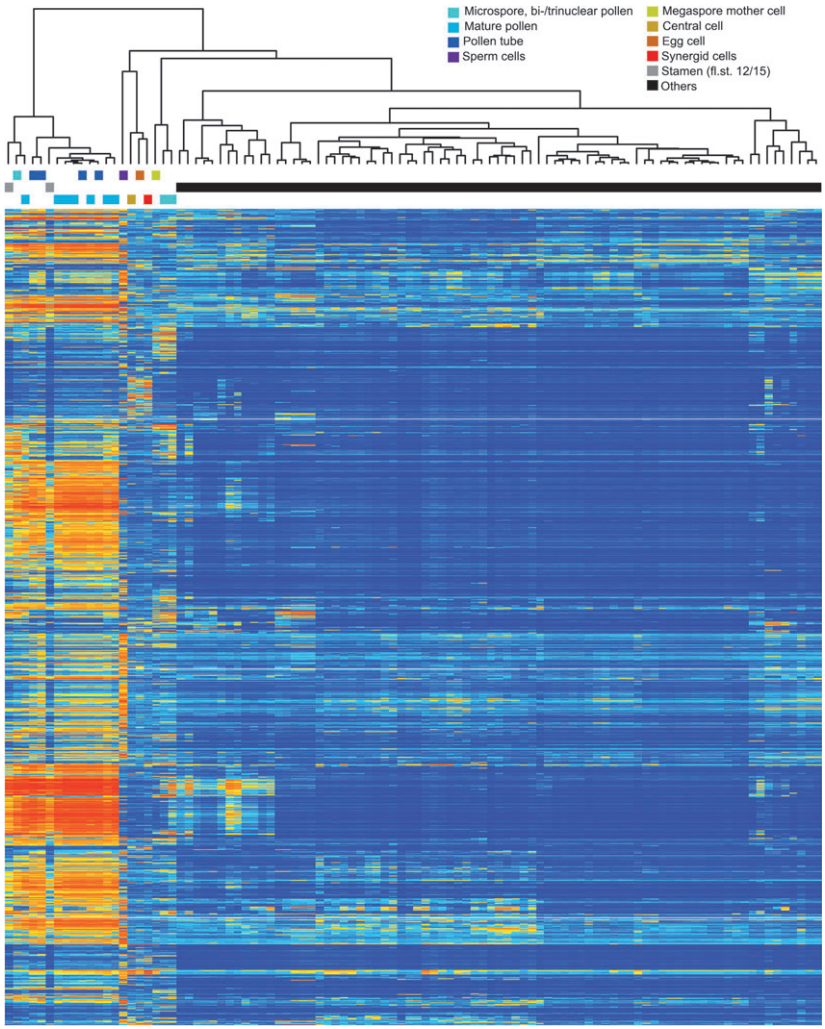


Figure 3. Genes preferentially expressed in gametophytic tissues and cell types.
Expression values (log_2 scale, calculated with robust multi-array analysis [RMA]; Irizarry *et al.*, 2005) of genes preferentially expressed in individual cells or tissue types from the male and female germline lineages are summarized in a heatmap (blue/red indicate low/high expression values). Replicates are averaged. The data set consisted of several mixed tissues, and specific tissue and cell types from Birnbaum *et al.* (2003), Honys and Twell (2004), Nawy *et al.* (2005), Pina *et al.* (2005), Schmid *et al.* (2005), Yu *et al.* (2005), Lee *et al.* (2006), Levesque *et al.* (2006), Brady *et al.* (2007), Borges *et al.* (2008), Wang *et al.* (2008), Qin *et al.* (2009), Yadav *et al.* (2009), Wuest *et al.* (2010), Schmidt *et al.* (2011). Others comprise sporophytic tissues such as siliques, rosette leaves, cotyledons, roots, root xylem, inflorescences, seeds, etc. Data were processed as described in Schmidt *et al.* (2011), except using an updated annotation of the ATH1 microarray (brainarray.mbni.med.umich.edu, TAIRG, version 14), an adjusted *P*-value cutoff of 0.05 and a minimal fold-change of four (on log_2 scale). Abbreviation: fl. st., floral stage.

Arabidopsis pollen together with Affymetrix 8K GeneCHIPs, representing approximately 8000 of the currently 33 602 annotated loci (TAIR10; http://www.arabidopsis.org/) (Table 1). In 2003, two independent studies identified 1584 and 992 genes expressed in pollen using this microarray, respectively (Table 1, Figure 2; Becker *et al.*, 2003; Honys and Twell, 2003). Only 1 year later, studying the expression in uninucleate microspores, binucleate, trinucleate and mature pollen, a total of 13 977 male gametophyte expressed genes were identified using the Affymetrix ATH1 array (Honys and Twell, 2004). On this array more than 22 500 probesets are spotted, originally designed for the detection of approximately 24 000 genes (http://www.affymetrix.com/). It was estimated that 61.9% of all genes represented on the Affymetrix ATH1 array are expressed in the male gametophyte (Honys and Twell, 2004). Using the Affymetrix ATH1 array, on average 6044 genes (Figure 2; 7235, 6587, 5004, 7177, 3954 and 6304, respectively) were identified to be expressed in mature pollen in independent studies (Table 1, Figure 2; Honys and Twell, 2004; Pina *et al.*, 2005; Schmid *et al.*, 2005; Borges *et al.*, 2008; Wang *et al.*, 2008; Qin *et al.*, 2009). These numbers demonstrate that the technical advance in array technology from the Affymetrix 8K GeneCHIP to the Affymetrix ATH1 array led to the identification of, on average, >4.5 times more expressed genes. Nevertheless, the highest and lowest estimates for expression in

Arabidopsis mature pollen differ by 3281 genes. These differences are probably due to different pollen harvesting methods, different Arabidopsis accessions, and different algorithms used for decision on presence or absence of expression. Microarrays were also used for transcriptional profiling of mature pollen from maize, soybean (*Glycine max*) and *Petunia axillaries*, and uninucleate microspores, bicellular and tricellular pollen from rice (Table 1, Figure 2; Ma *et al.*, 2008; Haerizadeh *et al.*, 2009; Ishimizu *et al.*, 2010; Wei *et al.*, 2010). Apart from this, genes expressed in the generative cell from *L. longiflorum* have been identified using cDNA microarrays (Table 1; Okada *et al.*, 2007). The transcriptome of isolated sperm cells from Arabidopsis was determined using Affymetrix ATH1 array, leading to the identification of 5829 sperm-cell-expressed genes (Table 1, Figure 2; Borges *et al.*, 2008). Using morphology-based selection and cDNA spotted microarrays, gene expression in the dimorphic sperm cells of *Plumbago zeylanica* was analysed separately (Table 1, Figure 2; Gou *et al.*, 2009).

Only recently, earlier developmental stages during microsporogenesis, i.e. Arabidopsis MiMCs, were studied to obtain new insights into the transcriptional basis of meiosis (Table 1, Figure 2; Chen *et al.*, 2010; Libeau *et al.*, 2011; Yang *et al.*, 2011). Using high-throughput sequencing of RNA (RNA-Seq) around 21 500 annotated loci likely to be expressed were identified (Yang *et al.*, 2011, 19 829 with at least one read in both replicates; Chen *et al.*, 2010, 23 843 with at least one read per million reads). A direct comparison with the data from Libeau *et al.* (2011), where the transcriptome of MiMCs was measured using the Complete Arabidopsis Transcriptome MicroArray (CATMA) is, however, difficult because the authors do not provide estimates of the transcriptome size. Nevertheless, the total number of genes found to be expressed in MiMCs using RNA-Seq is well above any previous reports from any studied cell type of the male germline lineage of Arabidopsis. Apart from the size of the transcriptome at this developmental stage, this largely reflects technical advances in the technology of transcriptome profiling by RNA-Seq as compared to microarrays, as will be discussed in more detail below. The different developmental stages of the male germline – from pre-meiotic MiMCs to the tricellular pollen – were also analysed in rice using LAM and microarrays (Table 1, Figure 2; Hobo *et al.*, 2008; Hirano *et al.*, 2008; Suwabe *et al.*, 2008; Tang *et al.*, 2010). Consistently, the studies led to the estimation of the expression of approximately 60% or more of all genes in the genome in either rice MiMCs (17 196 of 29 008 genes with representative probes on the array; Tang *et al.*, 2010) or Arabidopsis MiMCs (studies using RNA-Seq, Chen *et al.*, 2010; Yang *et al.*, 2011). Even though the transcriptome size at different stages of male germline development cannot always be directly compared due to the use of different profiling and isolation techniques, a significant reduction in size and complexity of the transcriptome from microsporogenesis, over early stages of microgametogenesis, to mature pollen is evident. In addition, Arabidopsis mature pollen was characterized by a relatively small transcriptome size compared with that of vegetative tissues, consistent with findings from a study analysing the soybean pollen transcriptome (Figure 2; Honys and Twell, 2004; Pina *et al.*, 2005; Schmid *et al.*, 2005; Haerizadeh *et al.*, 2009). The number of expressed genes is amazingly close to the estimated 20 000 transcripts in *Tradescantia paludosa* pollen that was based on hybridization kinetics (Willing and Mascarenhas, 1984).

In contrast to the relatively well-studied male germline, the transcriptional networks underlying female gametophyte development have only recently been investigated. To date, the transcriptomes of cell types of the mature embryo sac from Arabidopsis (egg cell, central cell, synergids) and rice (egg cell, synergids), and the Arabidopsis MeMC (Table 1, Figure 2; Wuest *et al.*, 2010; Ohnishi *et al.*, 2011; Schmidt *et al.*, 2011) have been described. Using LAM in combination with Affymetrix ATH1 array, 9115 genes were identified with evidence of expression in the Arabidopsis MeMC (Schmidt *et al.*, 2011), only slightly more than the 8850 genes identified to be expressed in the cells of the mature gametophyte (7171 in the egg cell, 7287 in the central cell and 5628 in synergids; Wuest *et al.*, 2010). A direct comparison with the transcriptome sizes of rice egg cells and synergids is not possible as the authors did not provide such estimates (Ohnishi *et al.*, 2011). From these results, the complexity of the transcriptome is not reduced to a similar extent during female as during male germline development. However, more than twice the number of genes with evidence of expression in Arabidopsis central cells have been identified using LAM in combination with RNA-Seq than previously identified using LAM and the Affymetrix ATH1 array [17 419 (Schmid *et al.*, 2012) compared with 7287 (Wuest *et al.*, 2010) genes]. This suggests a superior performance of RNA-Seq in the detection of expressed genes as compared to the broadly used Affymetrix ATH1 array.

RNA-SEQ OUTPERFORMS MICROARRAYS IN TERMS OF DETECTION RANGE AND FOR TRANSCRIPTOME PROFILING OF NON-MODEL SPECIES

For Arabidopsis transcriptional profiling the Affymetrix ATH1 array is so far the most frequently used platform, offering the advantage that a high number of different cell and tissue types can be directly compared (Schmid *et al.*, 2005; Wuest *et al.*, 2010; Schmidt *et al.*, 2011). Nonetheless, microarrays have several limitations: (i) high background levels due to cross-hybridization, (ii) a lack of sensitivity at low and high expression levels, and (iii) reliance upon existing knowledge about the genome sequence (Wang *et al.*, 2009). In addition, some microarrays designed for direct transcriptional profiling (i.e. non-tiling arrays such as

the Affymetrix ATH1 or CATMA arrays) can become outdated in terms of transcriptome coverage (e.g. ATH1 and CATMA arrays cover only around 64 and 66% of the 33 602 annotated loci in TAIR10), and do not offer the possibility of detecting previously unknown transcribed regions, and splice or sequence variants. Apart from this, probes for detecting genes with preferential or specific expression in the gametophytes are under-represented on the Affymetrix ATH1 array as compared with probes for detection of genes preferentially expressed in sporophytic tissues (Jones-Rhoades *et al.*, 2007).

RNA-Seq has the potential to overcome these limitations (Marioni *et al.*, 2008; Wang *et al.*, 2009), and therefore also offers the opportunity to study organisms lacking reference sequences, or to identify novel loci and alternative splicing events (Trapnell *et al.*, 2010). In terms of transcriptome size, RNA-Seq detects far more expressed genes than any study using Affymetrix ATH1 arrays for the profiling of cell types from the male or female germline lineages (Chen *et al.*, 2010; Yang *et al.*, 2011; Schmid *et al.*, 2012). Beside the effect of whole genome coverage, the difference is probably due to the higher sensitivity of RNA-Seq, as many genes seem to be expressed at a level that is not distinguishable from the background on the Affymetrix ATH1 arrays (Yang *et al.*, 2011; Schmid *et al.*, 2012). Interestingly, the increase in transcriptome size was not proportional for all classes of genes but more strongly affected certain gene classes, which are likely to be important for developmental processes and specific cellular functions (Schmid *et al.*, 2012).

Another prominent feature in RNA-Seq data is the presence of reads aligning to non-exonic regions (7% and 16% of all uniquely aligning reads in Schmid *et al.*, 2012; Yang *et al.*, 2011, respectively), including introns, regions flanking annotated loci, and isolated intergenic regions. The high number of non-exonic alignments in the central cell compared with other RNA-Seq transcriptomes (7% in a pool of organs and seedlings, Filichkin *et al.*, 2010; 3.5% in unopened flower buds, Lister *et al.*, 2008) may indicate transcriptional alterations prevalent in the central cell and novel transcribed regions that are specific to this cell type (Schmid *et al.*, 2012).

RNA-Seq has also been used for transcriptional profiling in non-model organisms. One possible approach for data analysis is the alignment of the reads to known sequences from a closely related organism. Szövényi *et al.* (2011) chose this strategy to compare the sporophytic with the gametophytic generation of the water moss *Funaria hygrometrica*, using reference sequences from *Physcomitrella patens*. Around 30% of the reads could be aligned to regions with an average nucleotide similarity of 95% (range 77–100%), indicating close genetic relatedness of the two species (Szövényi *et al.*, 2011). However, this similarity estimate may be biased towards a high value considering that alignments in more diverse regions are likely to fail the alignment criteria (Szövényi *et al.*, 2011). A limiting factor of this approach is not only the availability of reference sequences from a closely related species but also the read length and the total number of reads. Given the need for a permissive alignment strategy, the approach may be feasible for experiments with a relatively small number of long reads (approximately 600 000 high-quality reads obtained with the 454 pyrosequencer used in Szövényi *et al.*, 2011; average length not provided by the authors), but may perform poorly in an experiment with millions of short reads. In this case, *de novo* assembly of short reads into transcripts may perform significantly better (example given in Schmid *et al.*, 2012). This approach has recently been used to characterize the transcriptome of the (homosporous) gametophyte of the bracken fern *Pteridium aquilinum*, which has a diploid chromosome count of $2n = 104$, and a genome size of about 9.8 Gbp (Der *et al.*, 2011). The authors detected around 52 000 unique sequences (unigenes) from which 62% showed high similarities to known proteins (NCBI non-redundant protein database, http://www.ncbi.nlm.nih.gov). Notably, the data not only represented an 865-fold increase over the EST data available prior to the study on GenBank, but also led to the identification of 548 potentially amplifiable simple sequence repeats (SSRs) that may be used for genotyping (Der *et al.*, 2011). In addition, homologues of more than 50% of the presumably gametophyte-specific genes from Arabidopsis were identified (Der *et al.*, 2011; the list of gametophyte-specific genes in Arabidopsis was based on the data from Honys and Twell, 2004; Yu *et al.*, 2005; Wuest *et al.*, 2010). This indicates that, in the long run, RNA-Seq used for non-model plants or plants without sequenced reference genome can provide important insights in the development and evolution of the germline lineage and the alternation of generations in land plants.

NOVEL INSIGHTS IN TRANSCRIPTIONAL BASIS UNDERLYING GERMLINE SPECIFICATION

Studies analysing the transcriptional basis underlying male germline determination, sperm cell fate and pollen development provided new insights into the molecular mechanisms governing these important reproductive processes. Consistently, during male germline development a trend to reduce transcriptome size and complexity over the course of microsporogenesis and microgametogenesis has been observed (Honys and Twell, 2004; Wei *et al.*, 2010). While ≥60% of genes have been estimated to be expressed at onset of male germline development in pre-meiotic MiMCs (Tang *et al.*, 2010; Yang *et al.*, 2011), the transcriptome size of mature pollen has been estimated to comprise ≤30% of annotated loci (Pina *et al.*, 2005; Schmid *et al.*, 2005). However, as different transcriptional profiling methods have been used in these studies, and genes preferentially expressed in gametophytes are less represented on the

Affymetrix ATH1 array used to estimate the pollen transcriptome size, this difference might be overestimated. Nevertheless, despite the reduction of the overall transcriptome size during pollen maturation, an increasing functional specification of genes expressed in pollen has been observed, leading to estimates of 10–26% of pollen-specific genes (Figure 3; reviewed by Borg *et al.*, 2009). Analysing enriched gene expression at each developmental stage from the uninucleate microspore, bicellular and tricellular pollen, to the mature pollen grain, Wei *et al.* (2010) described a 'U-type' change in the numbers of preferentially expressed genes per developmental stage in rice and Arabidopsis, reaching a maximum level in mature pollen grains. Consistently, a reduced diversity of transcripts together with a functional skew towards transcripts related to cytoskeletal, cell wall and signalling processes have been described for mature pollen, probably important for germination, pollen tube growth and double fertilization (Honys and Twell, 2003, 2004; Pina *et al.*, 2005; Schmid *et al.*, 2005; Becker and Feijó, 2007; Borg *et al.*, 2009; Haerizadeh *et al.*, 2009). In addition, transcripts for translation and transcription were underrepresented, with the exception of certain classes of transcription factors, markedly including non-classical MADS-box transcription factors, i.e. type I and MIKC* (Honys and Twell, 2004; Pina *et al.*, 2005; reviewed by Grennan, 2007; Borg *et al.*, 2009). Interestingly, together with the RWP-RK domain and reproductive meristem (REM) transcription factor families, type I MADS domain transcription factors have also been identified as being up-regulated in the female gametophyte in comparison with other tissues, and were found to be exclusively enriched in reproductive tissues (Wuest *et al.*, 2010). This is in good agreement with recent studies on the expression and role of type I MADS box proteins during reproductive development (Bemer *et al.*, 2010; reviewed by Masiero *et al.*, 2011). This suggests that transcriptional profiling and enrichment analyses can aid in the identification of genes crucial for – or specifically expressed during – distinct stages of germline development and reproduction (Figure 3).

While certain genes and functions might be shared between male and female reproductive lineages, others are clearly distinct (Figure 3). Interestingly, enriched expression of PAZ and PIWI domain-encoding proteins is a dominant feature of the egg transcriptome (Wuest *et al.*, 2010). While small RNA pathways were first thought to be absent in Arabidopsis pollen and have not been detected in soybean pollen (Pina *et al.*, 2005; Haerizadeh *et al.*, 2009), expression of genes involved in small RNA pathways has subsequently been detected in Arabidopsis pollen and sperm (Borges *et al.*, 2008; Grant-Downton *et al.*, 2009a). Expression of genes involved in small RNA pathways has also been observed during megasporogenesis (Schmidt *et al.*, 2011). However, expression patterns in the MeMC were distinct from those of male or female gametophytes and gametes (Schmidt *et al.*, 2011). In addition to studying the transcriptional profile of genes involved in small RNA pathways, expression of known and novel small RNAs in the male germline has also been investigated using RNA-Seq or miRCURY LNA microarrays (Table 2) (Chambers and Shuai, 2009; Grant-Downton *et al.*, 2009b; Wei *et al.*, 2011).

In contrast to the relatively high number of studies addressing the transcriptional basis of microgametogenesis, only a few recent studies analyse gene expression during microsporogenesis (Chen *et al.*, 2010; Tang *et al.*, 2010; Libeau *et al.*, 2011; Yang *et al.*, 2011). The transcriptomes of Arabidopsis MiMCs isolated by micromanipulation and of rice pre-meiotic MiMCs isolated by laser microdissection has recently been studied with the purpose of identifying new genes playing a role in meiosis or in the context of meiotic cell divisions (Chen *et al.*, 2010; Tang *et al.*, 2010; Libeau *et al.*, 2011; Yang *et al.*, 2011). Consistently, Yang *et al.* (2011) reported expression of all 71 genes with described functions in meiosis, while enrichment of a number of meiotic genes in Arabidopsis MiMCs has been reported in other studies (Chen *et al.*, 2010; Libeau *et al.*, 2011). Interestingly, Tang *et al.* (2010) identified pathways important for meiotic recombination and cell cycle progression as well as expression of known meiotic genes enriched in pre-meiotic MiMCs, in agreement with the hypothesis that the transcriptional basis relevant for meiosis is already set up before its onset. Also, in Arabidopsis MeMCs sampled predominantly before meiosis to prophase of meiosis I, a number of genes with documented functions in meiosis but not in somatic tissues were found to be expressed (Schmidt *et al.*, 2011). Importantly, this study documented the prevalence of the biological process translation as well as the relevance of ATP-dependent RNA helicases in MeMCs, which play a role in the specification of the female germline lineage. These regulatory features are shared by the plant and animal germline (Schmidt *et al.*, 2011). Similarly, expression of 89

Table 2 Recent studies addressing small RNAs during development of the male and female germline lineage

Developmental stage	Isolation technique	Species	Profiling method	Literature
Male germline lineage				
UNM, BCP, TCP	Percoll gradient centrifugation	*O. sativa* ssp. *japonica*	Solexa sequencing	Wei *et al.* (2011)
Mature pollen	Percoll gradient centrifugation	*A. thaliana*	454 sequencing	Grant-Downton *et al.* (2009b)
Mature pollen	Modified hand-held vacuum	*A. thaliana*	miRCURY LNA array	Chambers and Shuai (2009)

UNM, uninucleate microspore; BCP, bicellular pollen; TCP, tricellular pollen.

DEAD-box containing ATP-binding helicases has also been observed in MiMCs (Yang *et al.*, 2011). Together, recent studies addressing cell-type-specific profiling of distinct developmental stages during male and female germline development in angiosperms provided important insights in their underlying gene expression profiles, molecular functions and regulatory pathways. However, a more detailed discussion of these findings and pathways, for example with respect to hormone signalling, cell–cell communication or gene regulation, is outside the scope of this review.

CONCLUSIONS AND OUTLOOK

Over the last decade, methodological improvements in both cell- and tissue-type-specific isolation methods as well as rapidly evolving techniques for whole-genome transcriptional profiling have provided new insights into the transcriptional basis and molecular mechanisms underlying the specification and development of the plant germline. Within a few years, knowledge of genes expressed at certain developmental stages of the male or female germline lineage have increased by one to two orders of magnitude, allowing investigations of gene and pathway enrichment to identify the underlying molecular mechanisms. However, while a relatively high number of studies have addressed transcriptional profiles underlying the development of the male lineage, only a few studies have concentrated on the female lineage, due to the small number and inaccessibility of the cells involved. Nevertheless, these studies allowed the identification of major trends, like the distinctiveness of transcriptional patterns underlying male and female gametophyte development and the realization that similar genes and pathways are active during specification of the plant and animal germline (Wuest *et al.*, 2010; Schmidt *et al.*, 2011). As RNA-Seq allows investigations of almost all species of interest and is not restricted to an analysis of model systems with known and annotated genomes, it is foreseeable that in the next years these technological improvements will help us to gain a deeper understanding of plant germline development. In particular, broadening the investigations to non-model organisms spanning the phylogenetic tree of land plants is likely to yield exciting insights into the evolutionary trends with respect to the alternation of generations as well as the underlying molecular determinants of germline fate.

REFERENCES

Bai, X., Peirson, B.N., Dong, F., Xue, C. and Makaroff, C.A. (1999) Isolation and characterization of *SYN1*, a *RAD21*-like gene essential for meiosis in Arabidopsis. *Plant Cell*, **11**, 417–430.

Becker, J.D. and Feijó, J.A. (2007) How many genes are needed to make a pollen tube? Lessons from transcriptomics. *Ann. Bot.* **100**, 1117–1123.

Becker, J.D., Boavida, L.C., Carneiro, J., Haury, M. and Feijó, J.A. (2003) Transcriptional profiling of *Arabidopsis* tissues reveals the unique characteristics of the pollen transcriptome. *Plant Physiol.* **133**, 713–725.

Bemer, M., Heijmans, K., Airoldi, C., Davies, B. and Angenent, G.C. (2010) An atlas of type I MADS box gene expression during female gametophyte and seed development in *Arabidopsis*. *Plant Physiol.* **154**, 287–300.

Berger, F. and Twell, D. (2011) Germline specification and function in plants. *Annu. Rev. Plant Biol.* **62**, 461–484.

Bhatt, A.M., Lister, C., Page, T., Fransz, P., Findlay, K., Jones, G.H., Dickinson, H.G. and Dean, C. (1999) The *DIF1* gene of *Arabidopsis* is required for meiotic chromosome segregation and belongs to the *REC8/RAD21* cohesin gene family. *Plant J.* **19**, 463–472.

Birnbaum, K., Shasha, D.E., Wang, J.Y., Jung, J.W., Lambert, G.M., Galbraith, D.W. and Benfey, P.N. (2003) A gene expression map of the *Arabidopsis* root. *Science*, **302**, 1956–1960.

Boavida, L.C., Becker, J.D. and Feijó, J.A. (2005) The making of gametes in higher plants. *Int. J. Dev. Biol.* **49**, 595–614.

Borg, M., Brownfield, L. and Twell, D. (2009) Male gametophyte development: a molecular perspective. *J Exp Bot.* **60**, 1465–1478.

Borges, F., Gomes, G., Gardner, R., Moreno, N., McCormick, S., Feijó, J.A. and Becker, J.D. (2008) Comparative transcriptomics of *Arabidopsis* sperm cells. *Plant Physiol.* **148**, 1168–1181.

Brady, S.M., Orlando, D.A., Lee, J.Y., Wang, J.Y., Koch, J., Dinneny, J.R., Mace, D., Ohler, U. and Benfey, P.N. (2007) A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science*, **318**, 801–806.

Brukhin, V., Curtis, M. and Grossniklaus, U. (2005) The female gametophyte: no longer the forgotten generation. *Curr. Sci.* **89**, 1844–1852.

Casson, S., Spencer, M., Walker, K. and Lindsey, K. (2005) Laser capture microdissection for the analysis of gene expression during embryogenesis of *Arabidopsis*. *Plant J.* **42**, 111–123.

Chambers, C. and Shuai, B. (2009) Profiling microRNA expression in *Arabidopsis* pollen using microRNA array and real-time PCR. *BMC Plant Biol.* **9**, 87.

Chang, F., Wang, Y., Wang, S. and Ma, H. (2011) Molecular control of microsporogenesis in *Arabidopsis*. *Curr. Opin. Plant Biol.* **14**, 66–73.

Chen, C., Farmer, A.D., Langley, R.J., Mudge, J., Crow, J.A., May, G.D., Huntley, J., Smith, A.G. and Retzel, E.F. (2010) Meiosis-specific gene discovery in plants: RNA-Seq applied to isolated *Arabidopsis* male meiocytes. *BMC Plant Biol.* **10**, 280.

Day, R.C., Grossniklaus, U. and Macknight, R.C. (2005) Be more specific! Laser-assisted microdissection of plant cells. *Trends Plant Sci.* **10**, 397–406.

Deal, R.B. and Henikoff, S. (2011) The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana*. *Nat. Protoc.* **6**, 56–68.

Der, J.P., Barker, M.S., Wickett, N.J., dePamphilis, C.W. and Wolf, P.G. (2011) *De novo* characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. *BMC Genomics*, **12**, 99.

Dickinson, H.G. and Grant-Downton, R. (2009) Bridging the generation gap: flowering plant gametophytes and animal germlines reveal unexpected similarities. *Biol. Rev. Camb. Philos. Soc.* **84**, 589–615.

Dresselhaus, T., Lörz, H. and Kranz, E. (1994) Representative cDNA libraries from few plant cells. *Plant J.* **5**, 605–610.

Emmert-Buck, M.R., Bonner, R.F., Smith, P.D., Chuaqui, R.F., Zhuang, Z., Goldstein, S.R., Weiss, R.A. and Liotta, L.A. (1996) Laser capture microdissection. *Science*, **274**, 998–1001.

Engel, M.L., Chaboud, A., Dumas, C. and McCormick, S. (2003) Sperm cells of *Zea mays* have a complex complement of mRNAs. *Plant J.* **34**, 697–707.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.K. and Mockler, T.C. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* **20**, 45–58.

Gou, X., Yuan, T., Wei, X. and Russell, S.D. (2009) Gene expression in the dimorphic sperm cells of *Plumbago zeylanica*: transcript profiling, diversity, and relationship to cell type. *Plant J.* **60**, 33–47.

Grant-Downton, R., Hafidh, S., Twell, D. and Dickinson, H.G. (2009a) Small RNA pathways are present and functional in the angiosperm male gametophyte. *Mol. Plant* **2**, 500–512.

Grant-Downton, R., Le Trionnaire, G., Schmid, R., Rodriguez-Enriquez, J., Hafidh, S., Mehdi, S., Twell, D. and Dickinson, H. (2009b) MicroRNA and tasiRNA diversity in mature pollen of *Arabidopsis thaliana*. *BMC Genomics*, **10**, 643.

Grennan, A.K. (2007) An analysis of the *Arabidopsis* pollen transcriptome. *Plant Physiol.* **145**, 3–4.

Gross-Hardt, R., Kägi, C., Baumann, N., Moore, J.M., Baskar, R., Gagliano, W.B., Jürgens, G. and Grossniklaus, U. (2007) *LACHESIS* restricts gametic cell fate in the female gametophyte of *Arabidopsis*. *PLoS Biol.* **5**, e47.

Grossniklaus, U. (2011) Plant germline development: a tale of cross-talk, signaling, and cellular interactions. *Sex. Plant Reprod.* **24**, 91–95.

Grossniklaus, U., Moore, J.M. and Gagliano, W.B. (1998) Molecular and genetic approaches to understanding and engineering apomixis: *Arabidopsis* as a powerful tool. In *Advances in Hybrid Rice Technology*. Proceedings of the 3rd International Symposium on Hybrid Rice 1996 (Virmani, S.S., Siddiq, E.A. and Muralidharan, K., eds). Manila, Philippines: International Rice Research Institute, pp. 187–211.

Grossniklaus, U., Moore, J.M., Brukhin, V., Gheyselinck, J., Baskar, R., Vielle-Calzada, J.-P., Baroux, C., Page, D.R. and Spillane, C. (2002) Engineering of apomixis in crop plants: what can we learn from sexual model systems. In *Plant Biotechnology 2002 and Beyond* (Vasil, I.K., ed.). Dordrecht: Kluwer Academic Publishers, pp. 309–314.

Haerizadeh, F., Wong, C.E., Bhalla, P.L., Gresshoff, P.M. and Singh, M.B. (2009) Genomic expression profiling of mature soybean (*Glycine max*) pollen. *BMC Plant Biol.* **9**, 25.

Haig, D. and Wilczek, A. (2006) Sexual conflict and the alternation of haploid and diploid generations. *Philos. Trans. R. Soc. B* **361**, 335–343.

Hirano, K., Aya, K., Hobo, T., Sakakibara, H., Kojima, M., Shim, R.A., Hasegawa, Y., Ueguchi-Tanaka, M. and Matsuoka, M. (2008) Comprehensive transcriptome analysis of phytohormone biosynthesis and signaling genes in microspore/pollen and tapetum of rice. *Plant Cell Physiol.* **49**, 1429–1450.

Hobo, T., Suwabe, K., Aya, K. *et al.* (2008) Various spatiotemporal expression profiles of anther-expressed genes in rice. *Plant Cell Physiol.* **49**, 1417–1428.

Honys, D. and Twell, D. (2003) Comparative analysis of the *Arabidopsis* pollen transcriptome. *Plant Physiol.* **132**, 640–652.

Honys, D. and Twell, D. (2004) Transcriptome analysis of haploid male gametophyte development in *Arabidopsis*. *Genome Biol.* **5**, R85.

Hoshino, Y., Murata, N. and Shinoda, K. (2006) Isolation of individual egg cells and zygotes in *Alstroemeria* followed by manual selection with a micro-capillary-connected micropump. *Ann. Bot.* **97**, 1139–1144.

Hu, T.-X., Miao, Y.U. and Zhao, J. (2011) Techniques of cell type-specific transcriptome analysis and application in researches of sexual plant reproduction. *Front. Biol.* **6**, 31–39.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15.

Ishimizu, T., Kodama, H., Ando, T. and Watanabe, M. (2010) Molecular evidence that most RNAs required for germination and pollen tube growth are stored in the mature pollen grain in petunia. *Genes Genet. Syst.* **85**, 259–263.

Johnston, A.J., Meier, P., Gheyselinck, J., Wuest, S.E., Federer, M., Schlagenhauf, E., Becker, J.D. and Grossniklaus, U. (2007) Genetic substraction profiling identifies genes essential for *Arabidopsis* reproduction and reveals interaction between the female gametophyte and the maternal sporophyte. *Genome Biol.* **8**, R204.

Jones-Rhoades, M.W., Borevitz, J.O. and Preuss, D. (2007) Genome-wide expression profiling of the *Arabidopsis* female gametophyte identifies families of small, secreted proteins. *PLoS Genet.* **3**, 1848–1861.

Kerk, N.M., Ceserani, T., Tausta, S.L., Sussex, I.M. and Nelson, T.M. (2003) Laser capture microdissection of cells from plant tissues. *Plant Physiol.* **132**, 27–35.

Kumlehn, J., Kirik, V., Czihal, A., Altschmied, L., Matzk, F., Lörz, H. and Bäumlein, H. (2001) Parthenogenetic egg cells of wheat: cellular and molecular studies. *Sex. Plant Reprod.* **14**, 239–243.

Lê, Q., Gutièrrez-Marcos, J.F., Costa, L.M., Meyer, S., Dickinson, H.G., Lörz, H., Kranz, E. and Scholten, S. (2005) Construction and screening of substracted cDNA libraries from limited populations of plant cells: a comparative analysis of gene expression between maize egg cells and central cells. *Plant J.* **44**, 167–178.

Lee, J.Y., Colinas, J., Wang, J.Y., Mace, D., Ohler, U. and Benfey, P.N. (2006) Transcriptional and posttranscriptional regulation of transcription factor expression in *Arabidopsis* roots. *Proc. Natl Acad. Sci. USA*, **15**, 6055–6060.

Levesque, M.P., Vernoux, T., Busch, W. *et al.* (2006) Whole-genome analysis of the SHORT-ROOT developmental pathway in *Arabidopsis*. *PLoS Biol.* **4**, e143.

Libeau, P., Durandet, M., Granier, F., Marquis, C., Berthomé, R., Renou, J.P., Taconnat-Soubirou, L. and Horlow, C. (2011) Gene expression profiling of *Arabidopsis* meiocytes. *Plant Biol.* **13**, 784–793.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, **133**, 523–536.

Ma, J., Skibbe, D.S., Fernandes, J. and Walbot, V. (2008) Male reproductive development: gene expression profiling of maize anther and pollen ontogeny. *Genome Biol.* **9**, R181.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517.

Masiero, S., Columbo, L., Grini, P.E., Schnittger, A. and Kater, M.M. (2011) The emerging importance of type I MADS box transcription factors for plant reproduction. *Plant Cell*, **23**, 865–872.

Nawy, T., Lee, J.Y., Colinas, J., Wang, J.Y., Thongrod, S.C., Malamy, J.E., Birnbaum, K. and Benfey, P.N. (2005) Transcriptional profile of the *Arabidopsis* root quiescent center. *Plant Cell*, **17**, 1908–1925.

Nelson, T., Tausta, S.L., Gandotra, N. and Liu, T. (2006) Laser microdissection of plant tissue: what you see is what you get. *Annu. Rev. Plant Biol.* **57**, 181–201.

Ohnishi, T., Takanashi, H., Mogi, M., Takahashi, H., Kikuchi, S., Yano, K., Okamoto, T., Fujita, M., Kurata, N. and Tsutsumi, N. (2011) Distinct gene expression profiles in egg and synergid cells of rice as revealed by cell-type-specific microarrays. *Plant Physiol.* **155**, 881–891.

Okada, T., Bhalla, P.L. and Singh, M.B. (2006) Expressed sequence tag analysis of *Lilium longiflorum* generative cells. *Plant Cell Physiol.* **47**, 698–705.

Okada, T., Singh, M.B. and Bhalla, P.L. (2007) Transcriptome profiling of *Lilium longiflorum* generative cells by cDNA microarray. *Plant Cell Rep.* **26**, 1045–1052.

Pina, C., Pinto, F., Feijó, J.A. and Becker, J.D. (2005) Gene family analysis of the *Arabidopsis* pollen transcriptome reveals biological implications for cell growth, division control, and gene expression regulation. *Plant Physiol.* **138**, 744–756.

Qin, Y., Leydon, A.R., Manziello, A., Pandey, R., Mount, D., Denic, S., Vasic, B., Johnson, M.A. and Palanivelu, R. (2009) Penetration of the stigma and style elicits a novel transcriptome in pollen tubes, pointing to genes critical for growth in a pistil. *PLoS Genet.* **5**, e1000621.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* **37**, 501–506.

Schmid, M.W., Schmidt, A., Klostermeier, U.C., Barann, M., Rosenstiel, P. and Grossniklaus, U. (2012) A powerful method for transcriptional profiling of specific cells in eukaryotes: laser-assisted microdissection and RNA sequencing. *PLoS ONE*. **7**, e29685.

Schmidt, A., Wuest, S.E., Vijverberg, K., Baroux, C., Kleen, D. and Grossniklaus, U. (2011) Transcriptome analysis of the *Arabidopsis* megaspore mother cell uncovers the importance of RNA helicases for plant germline development. *PLoS Biol.* **9**, e1101155.

Sprunck, S. and Gross-Hardt, R. (2011) Nuclear behavior, cell polarity, and cell specification in the female gametophyte. *Sex. Plant Reprod.* **24**, 123–136.

Sprunck, S., Baumann, U., Edwards, K., Langridge, P. and Dresselhaus, T. (2005) The transcript composition of egg cells changes significantly following fertilization in wheat (*Triticum aestivum* L.). *Plant J.* **41**, 660–672.

Steffen, J.G., Kang, I.-H., Macfarlane, J. and Drews, G.N. (2007) Identification of genes expressed in the *Arabidopsis* female gametophyte. *Plant J.* **51**, 281–292.

Sundaresan, V., Springer, P., Volpe, T., Haward, S., Jones, J.D., Dean, C., Ma, H. and Martienssen, R. (1995) Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev.* **9**, 1797–1810.

Suwabe, K., Suzuki, G., Takahashi, H. *et al.* (2008) Separated transcriptomes of male gametophyte and tapetum in rice: validity of laser microdissection (LM) microarray. *Plant Cell Physiol.* **49**, 1407–1416.

Szövényi, P., Rensing, S.A., Lang, D., Wray, G.A. and Shaw, A.J. (2011) Generation-biased gene expression in a bryophyte model system. *Mol. Biol. Evol.* **28**, 803–812.

Takanashi, H., Ohnishi, T., Mogi, M., Okamoto, T., Arimura, S. and Tsutsumi, N. (2010) Studies of mitochondrial morphology and DNA amount in the rice egg cell. *Curr. Genet.* **56**, 33–41.

Tang, X., Zhang, Z.-Y., Zhang, W.-J., Zhao, X.-M., Li, X., Zhang, D., Liu, Q.-Q. and Tang, W.-H. (2010) Global gene profiling of laser-captured pollen mother cells indicates molecular pathways and gene subfamilies involved in rice meiosis. *Plant Physiol.* **154**, 1855–1870.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515.

Twell, D. (2011) Male gametogenesis and germline specification in flowering plants. *Sex. Plant Reprod.* **24**, 149–160.

Wang, Y., Zhang, W.-Z., Song, L.-F., Zou, J.-J., Su, Z. and Wu, W.-H. (2008) Transcriptome analysis show changes in gene expression to accompany pollen germination and tube growth in *Arabidopsis*. *Plant Physiol.* **148**, 1201–1211.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.

Wei, L.Q., Xu, W.Y., Deng, Z.Y., Su, Z., Xue, Y. and Wang, T. (2010) Genome-scale analysis and comparison of gene expression profiles in developing and germinated pollen in *Oryza sativa*. *BMC Genomics*, **11**, 338.

Wei, L.Q., Yang, L.F. and Wang, T. (2011) Deep sequencing on genome-wide scale reveals the unique composition and expression patterns of microRNAs in developing pollen of *Oryza sativa*. *Genome Biol.* **12**, R53.

Willing, R.P. and Mascarenhas, J.P. (1984) Analysis of the complexity and diversity of mRNAs from pollen and shoots of *Tradescantia*. *Plant Physiol.* **75**, 865–868.

Wuest, S.E., Vijverberg, K., Schmidt, A., Weiss, M., Gheyselinck, J., Lohr, M., Wellmer, F., Rahnenführer, J., von Merig, C. and Grossniklaus, U. (2010) *Arabidopsis* female gametophyte gene expression map reveals similarities between plant and animal gametes. *Curr. Biol.* **20**, 506–512.

Xin, H.P. and Sun, M.X. (2010) What we have learned from transcript profile analyses of male and female gametes in flowering plants. *Sci. China Life Sci.* **53**, 927–933.

Xin, H.P., Peng, X.B., Ning, J., Yan, T.T., Ma, L.G. and Sun, M.X. (2011) Expressed sequence-tag analysis of tobacco sperm cells reveals a unique transcriptional profile and selective persistence of paternal transcripts after fertilization. *Sex. Plant Reprod.* **24**, 37–46.

Xu, H., Weterings, K., Vriezen, W., Feron, R., Xue, Y., Derksen, J. and Mariani, C. (2002) Isolation and characterization of male-germ-cell transcripts in *Nicotiana tabacum*. *Sex. Plant Reprod.* **14**, 339–346.

Yadav, R.K., Girke, T., Pasala, S., Xie, M. and Reddy, G.V. (2009) Gene expression map of the *Arabidopsis* shoot apical meristem stem cell niche. *Proc. Natl Acad. Sci. USA*, **106**, 4941–4946.

Yadegari, R. and Drews, G.N. (2004) Female gametophyte development. *Plant Cell* **16**(Suppl), S133–S141.

Yang, W.-C., Ye, D., Xu, J. and Sundaresan, V. (1999) The *SPOROCYTELESS* gene of *Arabidopsis* is required for initiation of sporogenesis and encodes a novel nuclear protein. *Genes Dev.* **13**, 2108–2117.

Yang, H., Kaur, N., Kiriakopolos, S. and McCormick, S. (2006) EST generation and analysis towards identifying female gametophyte-specific genes in *Zea mays* L. *Planta*, **224**, 1004–1014.

Yang, H., Lu, P., Wang, Y. and Ma, H. (2011) The transcriptome landscape of *Arabidopsis* male meiocytes from high-throughput sequencing: the complexity and evolution of the meiotic process. *Plant J.* **65**, 503–516.

Yu, H.J., Hogan, P. and Sundaresan, V. (2005) Analysis of the female gametophyte transcriptome of *Arabidopsis* by comparative expression profiling. *Plant Physiol.* **139**, 1853–1859.

8.6 Cell-specific expression profiling of rare cell types as exemplified by its impact on our understanding of female gametophyte development

The following review is published in "Current Opinion in Plant Biology" (published by Elsevier Ltd, all rights reserved)[1]. I provided a draft for the section "Next generation sequencing technologies – an unbiased and flexible toolkit to allow comparative studies in non-model organisms" (except the last paragraph).

[1]Wuest, SE, Schmid, MW, and Grossniklaus, U (2013) Cell-specific expression profiling of rare cell types as exemplified by its impact on our understanding of female gametophyte development. Current Opinion in Plant Biology 16: 41–49.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Plant Biology

Cell-specific expression profiling of rare cell types as exemplified by its impact on our understanding of female gametophyte development

Samuel E Wuest[1,2], Marc W Schmid[2] and Ueli Grossniklaus[2]

Expression profiling of single cells can yield insights into cell specification, cellular differentiation processes, and cell type-specific responses to environmental stimuli. Recent work has established excellent tools to perform genome-wide expression studies of individual cell types, even if the cells of interest occur at low frequency within an organ. We review the advances and impact of gene expression studies of rare cell types, as exemplified by recently gained insights into the development and function of the angiosperm female gametophyte. The detailed transcriptional characterization of different stages during female gametophyte development has significantly helped to improve our understanding of cellular specification or cell–cell communication processes. Next-generation sequencing approaches — used increasingly for expression profiling — will now allow for comparative approaches that focus on agriculturally, ecologically or evolutionarily relevant aspects of plant reproduction.

Addresses
[1] Institute of Evolutionary Biology and Environmental Studies & Zürich-Basel Plant Science Center, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland
[2] Institute of Plant Biology & Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, CH-8008 Zürich, Switzerland

Corresponding author: Grossniklaus, Ueli (grossnik@botinst.uzh.ch)

Current Opinion in Plant Biology 2013, **16**:41–49

This review comes from a themed issue on **Growth and development**

Edited by **Michael Scanlon** and **Marja Timmermans**

For a complete overview see the Issue and the Editorial

Available online 29th December 2012

http://dx.doi.org/10.1016/j.pbi.2012.12.001

Introduction

Cell specification, cellular differentiation, and specific cellular responses to environmental stimuli involve changes in gene expression. Therefore, a view of the transcriptome of a cell provides a snapshot of the cellular instruction machinery that strongly depends on developmental stage and environmental inputs. Recent technological developments have enabled genome-wide expression experiments at reasonable costs [1]. In addition, cell type-specific transcriptional profiling has dramatically improved our understanding of biological processes (e.g. reviewed in [2]). Two major lessons have been learnt from the analysis of genome-wide expression data in individual cell types or specific tissues in plants.

First, the cellular context is important when studying developmental processes, because cell-specific gene expression is generally masked when performing studies at the organ level (reviewed in [3]). Ground-breaking novel insights into the development of plants have been made by profiling essentially all cell types that occur in the *Arabidopsis* root [4–7], through studying male gametophyte development (pollen) [8], or by expression profiling during *Arabidopsis* female gametophyte (embryo sac) development [9,10•,11••,12••] (see also below).

Second, multiple lines of evidence suggest that not only cellular differentiation processes are best understood at the single-cell level, but also responses of an organism to environmental stimuli: in genome-wide expression studies strong interactions between cellular identity and environmental variation have been observed, for example when examining the effects of stress or nutrient treatments on different cell types in the root [13–15].

Here, we summarize the tools that are available to isolate specific cell types from heterogeneous tissues as well as advances in transcriptional profiling methods. Additional reviews on related topics have been published recently [2,3,16,17], but here we briefly summarize and compare the tools with a focus on their suitability for cells that occur at low frequencies within a tissue. We also discuss what insights have been gained through cell-specific gene expression profiling of rare cell types in plants, as exemplified by studies on female reproductive processes.

Techniques used for the isolation of individual cell types

Initial approaches for genome-wide expression profiling largely focused on the profiling technologies themselves, meaning that experiments were often performed at the whole-plant, organ, or tissue level [18]. Only for certain cell types, such as those of the male gametophyte, could specific stages be collected and profiled relatively easily [8]. In recent years however, the scientific community has been creative in generating a variety of methods to isolate and profile distinct cell types at specific developmental stages. Here, we summarize recently applied approaches — and identify limits, strengths and weaknesses of these — with a special focus on their

Table 1

Summary of popular cell isolation methods developed in recent years

Cell isolation method	Limits of relative cell occurrence	Relative enrichment scores of cells[b]	Technique/costs	Further applications	Use in non-model organisms?
Biochemical isolation	N/A [20,21]	N/A	Cheap	Many	Yes, but limited to few, selected cell types
FACS	~1% (down to 0.1% if highly optimized)	~9–60× [44••,89]	Expensive equipment, extensive protocol	Proteomics, metabolomics, genome-wide chromatin structure	Relies on transgenic lines
LAM	<0.1%	Depends on morphology (up to ~10,000×) [9]	Expensive equipment, long protocol	For rare cells: mostly limited to expression profiling (low throughput) but has also been used for DNA methylation analyses	Yes
INTACT/INTACT-derived	<1–10%[a]	100–170× [44••]/up to 10,000× [47••]	Easy protocol, cheap	Genome-wide chromatin structure	Relies on transgenic lines
Micromanipulation	<0.1%	Depends on accessibility of cells (~up to 10,000×) [48]	Technically challenging, depends on morphology of cell type	For rare cells: limited to expression profiling (due to low throughput)	Yes

[a] The limits of the method have not been tested thoroughly.
[b] Relative enrichment scores are defined here as [no. of target cells in output/no. of non-target cells in output]/[no. target cells in input/no. of non-target cells in input], e.g. from 10% relative fraction in input to 99% relative fraction in output: enrichment score = (99/1)/(10/90) = 891.

application to the collection of rare cell types (Table 1). For the use of methods in applications using frequent cell types, we refer to previous recent reviews [2,3,19].

Biochemical purification of selected cell types

It is possible to isolate certain cell types using mechanical and/or biochemical enrichment procedures. This approach is only applicable to selected cell types, for example, guard cells [20], trichomes [21] and sperm cells (MA Schauer, Protein dynamics of pollen development, PhD thesis, University of Zürich, 2010), with specifically designed methods for each. Since it is possible to isolate relatively large numbers of cells using such procedures, they are not only suitable for transcriptomics [20,21], but also for proteomics [22–24, MA Schauer, Protein dynamics of pollen development, PhD thesis, University of Zürich, 2010], metabolomics [25], analyses of DNA methylation [26], or the determination of cell wall composition [21].

Genetic subtraction methods

Hereby, a genetic background that alters developmental processes or cell type abundance in a tissue is employed, for instance mutants with an increased number of stomata [27], mutants with altered floral organs [28], floral mutants in combination with inducible transgenic constructs [29,30], or mutants missing the female gametophyte [31–34]. These tools are fairly limited to the biological system under study and often rely on specific mutants and/or transgenic backgrounds that may be difficult and time-consuming to establish. Furthermore, the approach is subtractive and, by definition, genes that are expressed both in the surrounding tissue and the target cells cannot be detected.

Fluorescence-activated sorting of cells or nuclei

The approach relies on automatically sorting cells [4,7,35] or nuclei [36] that are tagged with a fluorescent marker (such as the green fluorescent protein) using a flow cytometry system. The method results in high yields and good enrichments, and is suitable for transcriptomics [4,7,35,37], proteomics [38], and metabolomics [3]. However, the method is usually limited to cells that have a relative occurrence of more than 0.5% (0.1% under highly optimized conditions) within the harvested tissue (*Kenneth Birnbaum, personal communication*). This requirement impedes, for example, the isolation of female gametophytic cells from carpels or even from isolated ovules.

Cell-specific tagging of RNA, RNA-binding proteins, or components of the ribosome

In several experiments, specific tagging and pull-down of either RNA (e.g. in *Drosophila* [39]), RNA-binding proteins, or ribosomal components [40–43] has yielded insights into cell-specific processes. For example, in a study of cell-specific responses to environmental variation, it was shown how hypoxic stress affects the translatome (i.e. the mRNA population associated with the ribosome in the process of translation) [43]. However, these methods have so far been used for more frequently occurring cell types only, and their use for rare cell types may not feasible [39].

Affinity-isolation of nuclei from specific cell types

A recently established, elegant method for cell-specific expression profiling is INTACT, the Isolation of Nuclei TAgged in specific Cell Types [44••,45]. The method relies on affinity purification of nuclei that carry a biotinylated fusion protein in their nuclear envelopes. Thus, it supersedes the need to isolate whole cells, and offers a cheap and simple alternative to performing expression profiling on specific cell types. It has been used in model organisms such as *Arabidopsis* or *Caenorhabditis elegans* [46]. A further development of the method yielded excellent results for isolating subpopulations of *Drosophila* brain cells [47••]. In the latter work, enrichments of around 95% were achieved from tagged nuclei in neuronal cells that only make up around 0.14% of the *Drosophila* brain. Such a technique would make it possible to work with even some of the rarest cell types in *Arabidopsis*, and also allow the study of chromatin features in these cells. The technology is still being optimized and future work will reveal whether it is useful for the study of megagametogenesis. Like fluorescence activated cell sorting the INTACT system relies on genetic transformation and the use of suitable promoters that are active only in certain cell types.

Microsurgical manipulation

This method relies on the use of micropipettes to isolate single cells dissociated from heterogeneous tissues. It has been applied to the isolation of various rare cell types, such as female gametophytic cells from diverse plant species (see also below) [48–50]. The method has a relatively low yield but can be used to isolate rare cell types as long as they can be distinguished from the surrounding cells. Although it is easier to recognize the cells as they are marked, this method does not absolutely require the use of transgenic lines and is therefore applicable to cell types for which no specific promoters are available, as well as to non-model organisms that cannot easily be transformed.

Laser-assisted microdissection (LAM)

This approach involves the isolation and extraction of specific tissue compartments [51–58], or even individual cell types [9,10•,11••], with the use of a laser. Hereby, a tissue containing the cells of interest is chemically fixed, embedded in paraffin or resin, and sectioned on a microtome. Cells are then isolated based on morphology and cytology. This allows the isolation of rare cell types within complex tissues and even the isolation of subcellular domains, for example, specific regions within syncytial structures such as the endosperm [55] or embryo sac (Schmid and Grossniklaus, unpublished). However, yields are often limited and the use of LAM is usually restricted to transcriptomics, although it has also been used for the analysis of cell type-specific DNA methylation patterns [26,59]. This method does not require the generation of transgenic lines but the cells to be isolated have to be distinguishable in a section.

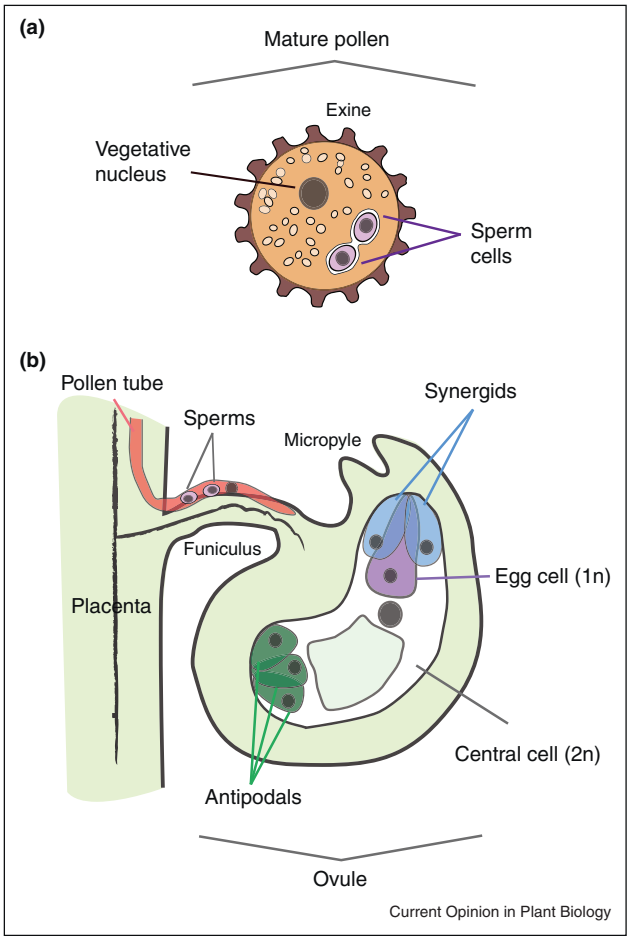
Cell-specific expression profiling — insights into female gametophyte development

The gametophytic generation in higher plants is of great agricultural importance. For instance, male-sterile plant lines can be generated for use in hybrid seed production through genetic or transgenic manipulations [60] or, in the future, apomictic crop species could be developed that would revolutionize breeding efforts [61,62]. However, a better understanding of the molecular basis of gametophyte development, which would open avenues for its manipulation for uses in crop breeding, has been hampered by its small size and, in particular for the female gametophyte, its inaccessibility. In the course of evolution, the gametophytic generation has become more and more restricted, such that in angiosperms it is represented as highly reduced organism, consisting of a few cells only, which is dependent on the sporophyte. Typically, at maturity male and female angiosperm gametophytes consist of only three and seven cells, respectively (Figure 1) [63].

Because of its simple organization and polar structure, the female gametophyte is considered an excellent system to study fundamental developmental processes, such as pattern formation, cell specification, and cell–cell communication [64–66]. However, the embryo sac of angiosperms is a very small structure, deeply embedded in sporophytic tissues, making it inaccessible for expression profiling studies. Classical genetic approaches have advanced our view of the molecular bases of developmental processes during gametogenesis [65,67]. Despite its success, the genetic approach has its shortcomings and only a few of the identified genes have been characterized in detail. To get a comprehensive overview over the genes involved in female gametophyte development it is thus desirable to use the advantages of current gene expression profiling tools to study embryo sac development.

The earliest gene expression studies used female gametophytic cells obtained by micromanipulation and subsequent EST-sequencing, first in maize [48] and later also in tobacco [68], wheat [69], and *Torrenia fournieri* [50]. Important insights into the communication between male and female gametophytes [49,50] and cellular specification [70] were only made possible through these EST-sequencing projects. For example, a recent EST-sequencing study in wheat revealed that RWP-RK domain-containing transcription factors exhibit egg cell-specific expression [71•]. The study also showed that two *Arabidopsis* homologues, *AtRKD1* and *AtRKD2*, could induce an egg cell-like transcriptional program when ectopically expressed in sporophytic tissue. Interestingly, the RKD transcription factors show similarity to the *MINUS DOMINANCE* gene of the green algae *Chlamydomonas reinhardtii*, which is important for sex-determination [72]. Even though loss-of-function phenotypes in *Arabidopsis* have not been described for

Figure 1



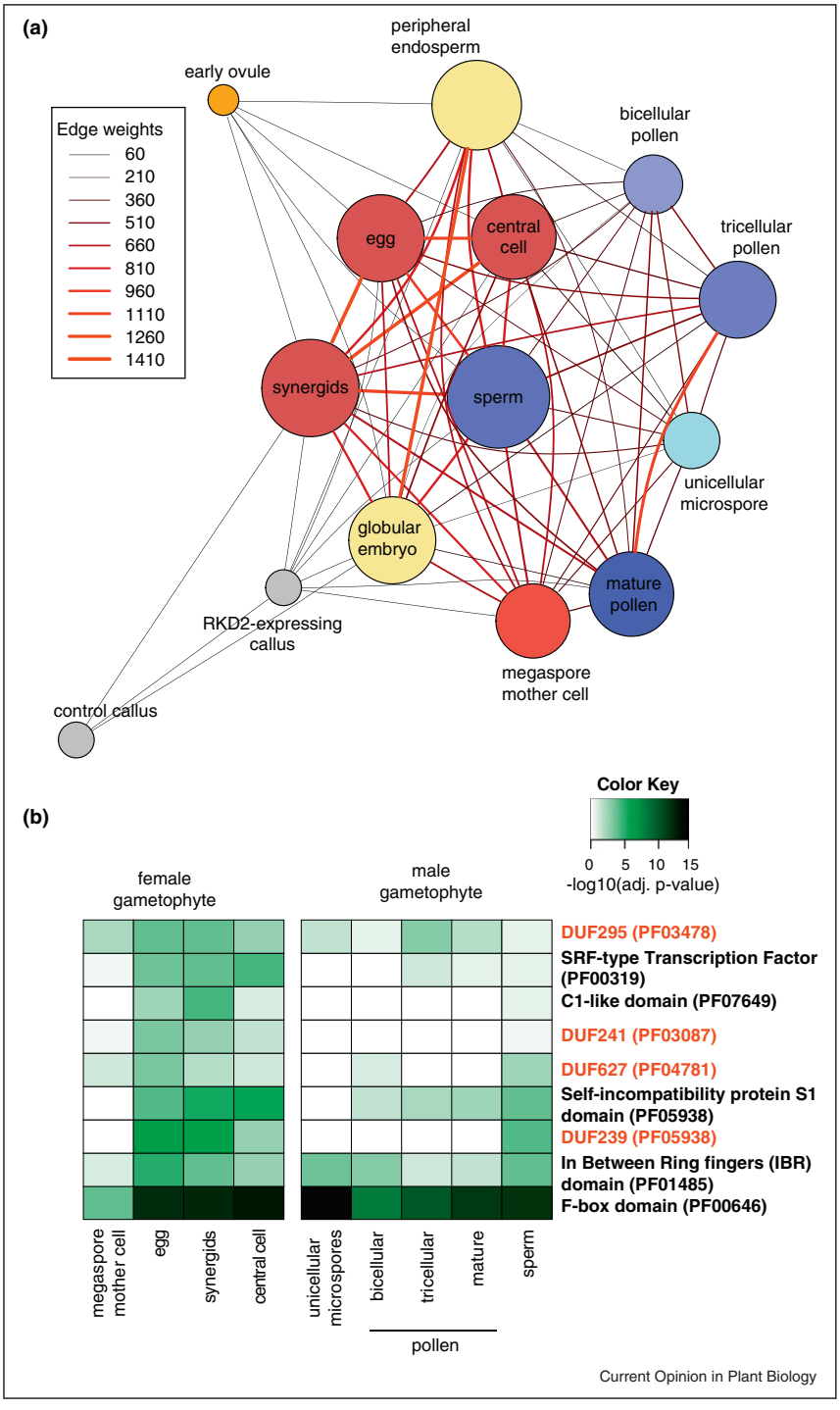
Schematic representations of the mature male and female gametophyte. **(a)** The male gametophyte, the pollen, develops mitotically from the microspores formed in the male reproductive structure, the anther. It consists of a vegetative cell and two sperm cells, the latter of which fertilize the two female gametes (central cell and egg) during double fertilization. **(b)** The female gametophyte, also termed embryo sac, is a seven-celled structure that develops within the sporophytic tissue of the ovule. It is made up by four different cell types, namely three antipodals (that can degenerate during the final stages of megagametogenesis), two synergids (accessory cells critical to the fertilization process), a homodiploid central cell which is fertilized by one sperm cell to give rise to the triploid endosperm, and an egg cell that is fertilized by the second sperm to give rise to the zygote.

AtRKD1 or *AtRKD2* because of genetic redundancy, it is likely that their gene products are involved in the specification of the female gametes as indicated by their ability to reprogram the transcriptional profile of sporophytic cells towards an egg cell fate [71•].

Extensions to these EST-sequencing projects include recent efforts that use a combination of single cell isolation and microarray profiling [9,11••,12••,73•] or the sequencing of RNA based on next generation sequencing methods (RNA-Seq — see also next paragraph) [10•]. These approaches have provided a more comprehensive view of the transcriptome and its dynamics during megagametogenesis. For example, profiles of an extensive selection of reproductive cell types in *Arabidopsis* as measured by the Affymetrix ATH1 GeneCHIP are available (Figure 2a). These data are an important resource to the research community, for example for reverse genetics [11••,12••] or to support the map-based cloning of mutant loci [74,75]. These studies also revealed that the transcriptomes of gametophytic cells contain a lot of uncharacterized genes and gene families (Figure 2b), some of which have most likely been missed during the development of the most popular microarray platforms [10•] (see also below). Thus, gametophytic cells express a large fraction of genes for which no function is known and their genetic characterization will reveal their roles during gametogenesis. Of course, such a study can be hampered by genetic redundancy, as was recently revealed for an egg cell-specific family of secreted peptides. Only mutation or downregulation of all five copies of this gene family revealed the specific function of these peptides in the fertilization process [76••].

Microgenomics approaches can be tied in with a functional framework on the mechanisms underlying development. A nice example of how transcriptional profiling can enhance our understanding of these — and resolve apparent discrepancies between different experimental observations — was recently published [12••]: it had been suggested that small RNA pathways are important for female gametophyte development and germline specification. The female germline is initiated with the determination of the megaspore mother cell (MMC) and terminates with the differentiation of the egg in the mature embryo sac [77]. In the first LAM-based gene expression map of the mature female gametophyte in *Arabidopsis* [9], it was found that genes of the *ARGONAUTE* (*AGO*) family, associated with small RNA pathways, are strongly expressed in the egg cell, suggesting an important role in the female gametophyte. On the other hand, mutations in the *AGO9* gene had been reported to result in the ectopic formation of MMCs [78••], suggesting that small RNA pathways are important in opposing female germline specification in somatic tissues. Thus, it was unclear whether small RNA pathways were involved in promoting or opposing megagametophyte development. Recently, applying LAM to identify genes expressed in the nucellar region of the ovule that contains developing megaspores, *AGO5* was found to show elevated expression [12••]. Indeed, a semidominant mutant allele of *AGO5* — which is an effector of small RNA pathways — results in a failure to initiate megagametogenesis. Similar defects were observed by expressing inhibitors of small RNA pathways in the nucellar region of the ovule. Therefore, there are apparently two opposing small RNA pathways that restrict MMC specification

Figure 2



Gene-sharing network of reproductive cell types in *Arabidopsis* as determined by a collection of experiments using the Affymetrix ATH1 GeneCHIP. **(a)** Gene-sharing network [79] showing a selection of reproductive cell type transcriptomes. Edge weights denote the number of genes shared between two nodes. Node size is proportional to the number of genes that are specifically enriched in a given cell type. **(b)** Selected PFAM-domains that are over-represented amongst gene products specifically expressed in different reproductive tissue (as determined in a). Color scale denotes *p*-values (with darker colors denoting smaller *p*-values), domain of unknown functions are highlighted in red.

to a single cell in the ovule and promote the development of the female gametophyte, respectively [12••,78••].

Next generation sequencing technologies — an unbiased and flexible toolkit to allow comparative studies in non-model organisms

In the past 10 years, microarrays have proven to be an indispensable tool for transcriptional profiling of rare cell types. The growing collection of publicly available data (e.g. based on the Affymetrix *Arabidopsis* ATH1 GeneCHIP), offers the opportunity to not only describe a newly obtained transcriptome, but also to identify cell type-specific expression patterns – the essence of cell differentiation [79]. Nonetheless, microarrays are constrained by the underlying technology and their design: firstly, hybridization-based measurements can exhibit high background levels due to cross-hybridization, and generally lack sensitivity at low and high expression levels; and secondly, the design relies upon existing knowledge of the genome sequence [1]. The latter is of particular importance for transcriptome arrays, which can become outdated regarding transcriptome coverage. For example, the ATH1 array lacks probes for ~36% of all genes, pseudo-genes, and transposable element genes annotated in TAIR10 [10•]. Intriguingly, a considerable fraction of these is likely to be important for reproductive development [10•,31,34].

An alternative that overcomes these limitations is RNA-Seq [1], which has been shown to be sensitive in detecting transcripts in rare cell types [10•,80,81]. In contrast to most microarray platforms, it also allows for a less biased analysis of the transcriptome and enables the identification of previously unannotated loci or transcript variants [82]. For example, a two to three times higher fraction of reads aligning uniquely to introns, regions flanking annotated loci, and isolated intergenic regions were observed in the central cell of *Arabidopsis* as compared to other RNA-Seq transcriptomes (16%, 7%, 7%, and 3.5% in central cells [10•], male meiocytes [81], pool of organs and seedlings [83], and unopened flower buds [84], respectively). This likely indicates novel transcribed regions and transcript variants that are specific to the central cell [10•].

RNA-Seq also offers the opportunity to study organisms that lack reference sequences. It may therefore promote the use of non-model species to study diverse ecologically, evolutionarily, and agriculturally relevant plant traits (reviewed in [85]). Depending on the organism and the availability of sequence information from closely related species, the analysis strategy may either be to firstly, align the reads to the reference sequences of a related species and use the annotation directly [86]; secondly, use those alignments for a reference-guided assembly [82]; or thirdly, perform a *de novo* assembly [87]. In many cases single cell specificity in non-model organisms may only be achieved with isolation methods that do not rely on the generation of transgenic plants, for example with microdissection or LAM (see example in [88]). For our understanding of the gametophytic generation in higher plants an extension of expression profiling methods to non-model organisms bears great promise for breakthroughs. For instance, recent efforts of characterizing the molecular bases of apomixis, the asexual reproduction through seeds, are increasingly focused on expression studies for a comparison of asexual with sexual reproduction [61]. Another example is the comparative approach taken to reveal the transcriptomes of the gametophytic and sporophytic generations in the water moss *Funaria hygrometrica*. This comparison revealed a weaker differentiation in gene expression between the two generations compared to *Arabidopsis*, which was attributed to the fact that both generations of *F. hygrometrica* are well developed and consist of a large number of cells, whereas *Arabidopsis* has dramatically reduced gametophytes [86].

Expression studies in individual cell types and under selected conditions have already revealed exciting insights into plant development and cell type-specific responses to environmental stimuli. A cell type-resolved view of gene expression in several plant species is now available, and will allow for comparative studies that will shed new light onto the evolution of developmental processes, including that of sexual and asexual reproduction.

Acknowledgements

The authors would like to acknowledge Chloé Wuest for reading of the manuscript, and Song Li (Duke University) for R scripts and helpful discussions. S.E.W. is supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme, U.G. receives funding from the Swiss National Science Foundation, the University Research Priority Program in Functional Genomics/Systems Biology, and COST-Action FA0903 through a grant of the 'Staatssekretariat für Bildung und Forschung' to support research in this field.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**:57-63.
2. Taylor-Teeples M, Ron M, Brady SM: **Novel biological insights revealed from cell type-specific expression profiling**. *Curr Opin Plant Biol* 2011, **14**:601-607.
3. Rogers ED, Jackson T, Moussaieff A, Aharoni A, Benfey PN: **Cell type-specific transcriptional profiling: implications for metabolite profiling**. *Plant J* 2012, **70**:5-17.
4. Brady SM, Orlando DA, Lee JY, Wang JY, Koch J, Dinneny JR, Mace D, Ohler U, Benfey PN: **A high-resolution root spatiotemporal map reveals dominant expression patterns**. *Science* 2007, **318**:801-806.
5. Moreno-Risueno MA, Van Norman JM, Moreno A, Zhang J, Ahnert SE, Benfey PN: **Oscillating gene expression determines competence for periodic *Arabidopsis* root branching**. *Science* 2010, **329**:1306-1311.

6. Iyer-Pascuzzi AS, Jackson T, Cui H, Petricka JJ, Busch W, Tsukagoshi H, Benfey PN: **Cell identity regulators link development and stress responses in the *Arabidopsis* root**. *Dev Cell* 2011, **21**:770-782

7. Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, Benfey PN: **A gene expression map of the *Arabidopsis* root**. *Science* 2003, **302**:1956-1960.

8. Honys D, Twell D: **Transcriptome analysis of haploid male gametophyte development in *Arabidopsis***. *Genome Biol* 2004, **5**:R85.

9. Wuest SE, Vijverberg K, Schmidt A, Weiss M, Gheyselinck J, Lohr M, Wellmer F, Rahnenfuhrer J, von Mering C, Grossniklaus U: ***Arabidopsis* female gametophyte gene expression map reveals similarities between plant and animal gametes**. *Curr Biol* 2010, **20**:506-512.

10. • Schmid MW, Schmidt A, Klostermeier UC, Barann M, Rosenstiel P, Grossniklaus U: **A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing**. *PLoS One* 2012, **7**:e29685.

This study describes the combination LAM and RNA-Seq for transcriptional profiling of rare cell types using the central cell of *Arabidopsis* as an example. Comparison to ATH1 microarray data reveals a largely improved sensitivity and accuracy. An exploration of a *de novo* assembly approach suggests that RNA-Seq can also be applied for the study of single cell transcriptomes in non-model organisms.

11. •• Schmidt A, Wuest SE, Vijverberg K, Baroux C, Kleen D, Grossniklaus U: **Transcriptome analysis of the *Arabidopsis* megaspore mother cell uncovers the importance of RNA helicases for plant germline development**. *PLoS Biol* 2011, **9**:e1001155.

The paper describes the meticulous sampling of megaspore mother cell from the *Arabidopsis* nucellar region by LAM and subsequent transcriptional profiling. Among a gene family of RNA helicases with strongly enriched expression in this cell type, the *MNEME* (*MEM*) gene was functionally characterized. *MEM* is shown to be involved in the restriction of the female germline to a single cell in the nucellus, such that *mem* mutations show aspects of apospory, an element of apomixis.

12. •• Tucker MR, Okada T, Hu Y, Scholefield A, Taylor JM, Koltunow AM: **Somatic small RNA pathways promote the mitotic events of megagametogenesis during female reproductive development in *Arabidopsis***. *Development* 2012, **139**:1399-1404.

The authors used LAM and microarray expression profiling of the nucellar region of developing *Arabidopsis* ovules. They describe elevated expression of Arabidopsis *AGO5* in this tissue, and functional analyses of small RNA pathway components suggest that these promote megagametogenesis. The finding complements the previously described role for certain small RNA pathways to restrict the female germline in the nucellus. The resulting model proposes that at least two small RNA pathways are active in ovules, one to restrict and the other to promote the reproductive potential of cells.

13. Gifford ML, Dean A, Gutierrez RA, Coruzzi GM, Birnbaum KD: **Cell-specific nitrogen responses mediate developmental plasticity**. *Proc Natl Acad Sci U S A* 2008, **105**:803-808.

14. Dinneny JR, Long TA, Wang JY, Jung JW, Mace D, Pointer S, Barron C, Brady SM, Schiefelbein J, Benfey PN: **Cell identity mediates the response of *Arabidopsis* roots to abiotic stress**. *Science* 2008, **320**:942-945.

15. Long TA, Tsukagoshi H, Busch W, Lahner B, Salt DE, Benfey PN: **The bHLH transcription factor POPEYE regulates response to iron deficiency in *Arabidopsis* roots**. *Plant Cell* 2010, **22**:2219-2236.

16. Pu L, Brady S: **Systems biology update: cell type-specific transcriptional regulatory networks**. *Plant Physiol* 2010, **152**:411-419.

17. Galbraith DW, Birnbaum K: **Global studies of cell type-specific gene expression in plants**. *Annu Rev Plant Biol* 2006, **57**:451-475.

18. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU: **A gene expression map of *Arabidopsis thaliana* development**. *Nat Genet* 2005, **37**:501-506.

19. Long TA: **Many needles in a haystack: cell-type specific abiotic stress responses**. *Curr Opin Plant Biol* 2011, **14**:325-331

20. Leonhardt N, Kwak JM, Robert N, Waner D, Leonhardt G, Schroeder JI: **Microarray expression analyses of *Arabidopsis* guard cells and isolation of a recessive abscisic acid hypersensitive protein phosphatase 2C mutant**. *Plant Cell* 2004, **16**:596-615.

21. Marks MD, Betancur L, Gilding E, Chen F, Bauer S, Wenger JP, Dixon RA, Haigler CH: **A new method for isolating large quantities of *Arabidopsis* trichomes for transcriptome, cell wall and other types of analyses**. *Plant J* 2008, **56**:483-492.

22. Amme S, Rutten T, Melzer M, Sonsmann G, Vissers JP, Schlesier B, Mock HP: **A proteome approach defines protective functions of tobacco leaf trichomes**. *Proteomics* 2005, **5**:2508-2518.

23. Zhao Z, Zhang W, Stanley BA, Assmann SM: **Functional proteomics of *Arabidopsis thaliana* guard cells uncovers new stomatal signaling pathways**. *Plant Cell* 2008, **20**:3210-3226.

24. Wu T, Wang Y, Guo D: **Investigation of glandular trichome proteins in *Artemisia annua* L. using comparative proteomics**. *PLoS One* 2012, **7**:e41822.

25. Ebert B, Zoller D, Erban A, Fehrle I, Hartmann J, Niehl A, Kopka J, Fisahn J: **Metabolic profiling of *Arabidopsis thaliana* epidermal cells**. *J Exp Bot* 2010, **61**:1321-1335.

26. Wöhrmann HJ, Gagliardini V, Raissig MT, Wehrle W, Arand J, Schmidt A, Tierling S, Page DR, Schob H, Walter J, Grossniklaus U: **Identification of a DNA methylation-independent imprinting control region at the *Arabidopsis MEDEA* locus**. *Genes Dev* 2012, **26**:1837-1850.

27. Bergmann DC, Lukowitz W, Somerville CR: **Stomatal development and pattern controlled by a MAPKK kinase**. *Science* 2004, **304**:1494-1497.

28. Wellmer F, Riechmann JL, Alves-Ferreira M, Meyerowitz EM: **Genome-wide analysis of spatial gene expression in *Arabidopsis* flowers**. *Plant Cell* 2004, **16**:1314-1326.

29. Gomez-Mena C, de Folter S, Costa MM, Angenent GC, Sablowski R: **Transcriptional program controlled by the floral homeotic gene *AGAMOUS* during early organogenesis**. *Development* 2005, **132**:429-438.

30. Wuest SE, O'Maoileidigh DS, Rae L, Kwasniewska K, Raganelli A, Hanczaryk K, Lohan AJ, Loftus B, Graciet E, Wellmer F: **Molecular basis for the specification of floral organs by *APETALA3* and *PISTILLATA***. *Proc Natl Acad Sci U S A* 2012, **109(33)**:13452-13457.

31. Sanchez-Leon N, Arteaga-Vazquez M, Alvarez-Mejia C, Mendiola-Soto J, Duran-Figueroa N, Rodriguez-Leal D, Rodriguez-Arevalo I, Garcia-Campayo V, Garcia-Aguilar M, Olmedo-Monfil V *et al.*: **Transcriptional analysis of the *Arabidopsis* ovule by massively parallel signature sequencing**. *J Exp Bot* 2012, **63**:3829-3842.

32. Yu HJ, Hogan P, Sundaresan V: **Analysis of the female gametophyte transcriptome of *Arabidopsis* by comparative expression profiling**. *Plant Physiol* 2005, **139**:1853-1869.

33. Johnston AJ, Meier P, Gheyselinck J, Wuest SE, Federer M, Schlagenhauf E, Becker JD, Grossniklaus U: **Genetic subtraction profiling identifies genes essential for *Arabidopsis* reproduction and reveals interaction between the female gametophyte and the maternal sporophyte**. *Genome Biol* 2007, **8**:R204.

34. Jones-Rhoades MW, Borevitz JO, Preuss D: **Genome-wide expression profiling of the *Arabidopsis* female gametophyte identifies families of small, secreted proteins**. *PLoS Genet* 2007, **3**:1848-1861.

35. Brady SM, Zhang L, Megraw M, Martinez NJ, Jiang E, Yi CS, Liu W, Zeng A, Taylor-Teeples M, Kim D *et al.*: **A stele-enriched gene regulatory network in the *Arabidopsis* root**. *Mol Syst Biol* 2011, **7**:459.

36. Zhang C, Barthelson RA, Lambert GM, Galbraith DW: **Global characterization of cell-specific gene expression through fluorescence-activated sorting of nuclei**. *Plant Physiol* 2008, **147**:30-40.

37. Nawy T, Lee JY, Colinas J, Wang JY, Thongrod SC, Malamy JE, Birnbaum K, Benfey PN: **Transcriptional profile of the *Arabidopsis* root quiescent center**. *Plant Cell* 2005, **17**:1908-1925

38. Petricka JJ, Schauer MA, Megraw M, Breakfield NW, Thompson JW, Georgiev S, Soderblom EJ, Ohler U, Moseley MA, Grossniklaus U, Benfey PN: **The protein expression landscape of the *Arabidopsis* root**. *Proc Natl Acad Sci U S A* 2012, **109**:6811-6818.

39. Miller MR, Robinson KJ, Cleary MD, Doe CQ: **TU-tagging: cell type-specific RNA isolation from intact complex tissues**. *Nat Methods* 2009, **6**:439-441.

40. Doyle JP, Dougherty JD, Heiman M, Schmidt EF, Stevens TR, Ma G, Bupp S, Shrestha P, Shah RD, Doughty ML *et al.*: **Application of a translational profiling approach for the comparative analysis of CNS cell types**. *Cell* 2008, **135**:749-762.

41. Heiman M, Schaefer A, Gong S, Peterson JD, Day M, Ramsey KE, Suarez-Farinas M, Schwarz C, Stephan DA, Surmeier DJ *et al.*: **A translational profiling approach for the molecular characterization of CNS cell types**. *Cell* 2008, **135**:738-748.

42. Jiao Y, Meyerowitz EM: **Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control**. *Mol Syst Biol* 2010, **6**:419.

43. Mustroph A, Zanetti ME, Jang CJ, Holtan HE, Repetti PP, Galbraith DW, Girke T, Bailey-Serres J: **Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in Arabidopsis**. *Proc Natl Acad Sci U S A* 2009, **106**:18843-18848.

44. •• Deal RB, Henikoff S: **A simple method for gene expression and chromatin profiling of individual cell types within a tissue**. *Dev Cell* 2010, **18**:1030-1040.

INTACT, a simple, cheap, and powerful technique to isolate cell-type specific nuclei is described in this paper. It relies on the epitope-tagging of proteins localized to the nuclear periphery and affinity-purification. The method is shown to have high sensitivity and good yields. It was used for expression as well as chromatin profiling of two root epidermal cell types.

45. Deal RB, Henikoff S: **The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana***. *Nat Protoc* 2011, **6**:56-68.

46. Steiner FA, Talbert PB, Kasinathan S, Deal RB, Henikoff S: **Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling**. *Genome Res* 2012, **22**:766-777.

47. •• Henry GL, Davis FP, Picard S, Eddy SR: **Cell type-specific genomics of *Drosophila* neurons**. *Nucleic Acids Res* 2012, **40**:9691-9704.

A modified INTACT method is described to isolate tagged nuclei from specific *Drosophila melanogaster* neurons. The authors managed to enrich extremely rare neuronal cell types with unprecedented recoveries and excellent yields for expression as well as chromatin profiling.

48. Dresselhaus T, Lorz H, Kranz E: **Representative cDNA libraries from few plant cells**. *Plant J* 1994, **5**:605-610.

49. Marton ML, Cordts S, Broadhvest J, Dresselhaus T: **Micropylar pollen tube guidance by *egg apparatus 1* of maize**. *Science* 2005, **307**:573-576.

50. Okuda S, Tsutsui H, Shiina K, Sprunck S, Takeuchi H, Yui R, Kasahara RD, Hamamura Y, Mizukami A, Susaki D *et al.*: **Defensin-like polypeptide LUREs are pollen tube attractants secreted from synergid cells**. *Nature* 2009, **458**:357-361.

51. Brooks L III, Strable J, Zhang X: **Microdissection of shoot meristem functional domains**. *PLoS Genet.* 2009, **5**:e1000476.

52. Cai S, Lashbrook CC: **Stamen abscission zone transcriptome profiling reveals new candidates for abscission control: enhanced retention of floral organs in transgenic plants overexpressing *Arabidopsis ZINC FINGER PROTEIN2***. *Plant Physiol* 2008, **146**:1305-1321.

53. Deeken R, Ache P, Kajahn I, Klinkenberg J, Bringmann G, Hedrich R: **Identification of *Arabidopsis thaliana* phloem RNAs provides a search criterion for phloem-based transcripts hidden in complex datasets of microarray experiments**. *Plant J* 2008, **55**:746-759

54. Kerk NM, Ceserani T, Tausta SL, Sussex IM, Nelson TM: **Laser capture microdissection of cells from plant tissues**. *Plant Physiol* 2003, **132**:27-35.

55. Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S *et al.*: **Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors**. *Proc Natl Acad Sci U S A* 2010, **107**:8063-8070.

56. Ohtsu K, Smith MB, Emrich SJ, Borsuk LA, Zhou R, Chen T, Zhang X, Timmermans MC, Beck J, Buckner B *et al.*: **Global gene expression analysis of the shoot apical meristem of maize (*Zea mays* L.)**. *Plant J* 2007, **52**:391-404.

57. Spencer MW, Casson SA, Lindsey K: **Transcriptional profiling of the *Arabidopsis* embryo**. *Plant Physiol* 2007, **143**:924-940.

58. Walia H, Josefsson C, Dilkes B, Kirkbride R, Harada J, Comai L: **Dosage-dependent deregulation of an *AGAMOUS-LIKE* gene cluster contributes to interspecific incompatibility**. *Curr Biol* 2009, **19**:1128-1132.

59. You W, Tyczewska A, Spencer M, Daxinger L, Schmid MW, Grossniklaus U, Simon SA, Meyers BC, Matzke AJ, Matzke M: **A typical DNA methylation of genes encoding cysteine-rich peptides in *Arabidopsis thaliana***. *BMC Plant Biol* 2012, **12**:51.

60. Dwivedi S, Perotti E, Ortiz R: **Towards molecular breeding of reproductive traits in cereal crops**. *Plant Biotechnol J* 2008, **6**:529-559.

61. Koltunow AM, Johnson SD, Okada T: **Apomixis in hawkweed: Mendel's experimental nemesis**. *J Exp Bot* 2011, **62**:1699-1707.

62. Spillane C, Curtis MD, Grossniklaus U: **Apomixis technology development-virgin births in farmers' fields?** *Nat Biotechnol* 2004, **22**:687-691.

63. Ma H, Sundaresan V: **Development of flowering plant gametophytes**. *Curr Top Dev Biol* 2010, **91**:379-412.

64. Grossniklaus U, Schneitz K: **The molecular and genetic basis of ovule and megagametophyte development**. *Semin Cell Dev Biol* 1998, **9**:227-238.

65. Sprunck S, Gross-Hardt R: **Nuclear behavior, cell polarity, and cell specification in the female gametophyte**. *Sex Plant Reprod* 2011, **24**:123-136.

66. Chevalier E, Loubert-Hudon A, Zimmerman EL, Matton DP: **Cell–cell communication and signalling pathways within the ovule: from its inception to fertilization**. *New Phytol* 2011, **192**:13-28.

67. Brukhin V, Curtis MD, Grossniklaus U: **The angiosperm female gametophyte: no longer the forgotten generation**. *Curr Sci* 2005, **89**:1844-1852.

68. Ning J, Peng XB, Qu LH, Xin HP, Yan TT, Sun M: **Differential gene expression in egg cells and zygotes suggests that the transcriptome is restructed before the first zygotic division in tobacco**. *FEBS Lett* 2006, **580**:1747-1752.

69. Kumlehn J, Kirik V, Czihal A, Altschmied L, Matzk F, Lorz H, Bäumlein H: **Parthenogenetic egg cells of wheat: cellular and molecular studies**. *Sex Plant Reprod* 2001, **14**:239-243.

70. Krohn NG, Lausser A, Juranic M, Dresselhaus T: **Egg cell signaling by the secreted peptide *ZmEAL1* controls antipodal cell fate**. *Dev Cell* 2012, **23**:219-225.

71. • Köszegi D, Johnston AJ, Rutten T, Czihal A, Altschmied L, Kumlehn J, Wüst SE, Kirioukhova O, Gheyselinck J, Grossniklaus U, Bäumlein H: **Members of the RKD transcription factor family induce an egg cell-like gene expression program**. *Plant J* 2011, **67**:280-291.

The authors describe the egg cell-specific expression of two RWP-RK domain containing transcription factors in wheat (*TaRKD1/2*), discovered by an EST sequencing approach. Similarly, two *Arabidopsis* homologues (*AtRKD1/2*) were found to be expressed specifically in egg cells. Ectopic expression of these two genes in *Arabidopsis* leads to proliferation of undifferentiated cells and the partial reprograming towards an egg cell-specific transcriptional program.

72. Ferris PJ, Goodenough UW: **Mating type in *Chlamydomonas* is specified by *mid*, the minus-dominance gene**. *Genetics* 1997, **146**:859-869

73. • Ohnishi T, Takanashi H, Mogi M, Takahashi H, Kikuchi S, Yano K, Okamoto T, Fujita M, Kurata N, Tsutsumi N: **Distinct gene expression profiles in egg and synergid cells of rice as revealed by cell type-specific microarrays**. *Plant Physiol* 2011, **155**:881-891.
This paper describes a thorough, genome-wide characterization of rice synergid and egg cell transcriptomes, representing the first whole-genome study on the egg apparatus in a monocot species.

74. Kessler SA, Shimosato-Asano H, Keinath NF, Wuest SE, Ingram G, Panstruga R, Grossniklaus U: **Conserved molecular components for pollen tube reception and fungal invasion**. *Science* 2010, **330**:968-971.

75. Leshem Y, Johnson C, Wuest SE, Song X, Ngo QA, Grossniklaus U, Sundaresan V: **Molecular characterization of the *glauce* mutant: a central cell-specific function is required for double fertilization in *Arabidopsis***. *Plant Cell* 2012, **24**:3264-3277.

76. •• Sprunck S, Rademacher S, Vogler F, Gheyselinck J, Grossniklaus U, Dresselhaus T: **Egg cell-secreted *EC1* triggers sperm cell activation during double fertilization**. *Science* 2012, **338**:1093-1097.
This paper demonstrates the role of the five redundant *EGG CELL1 (EC1)* genes in double fertilization. EC1 proteins accumulate in storage vesicles that exocytose upon sperm arrival. This leads to the redistribution of the potentially fusogenic protein HAPLESS2/GENERATIVE CELL SPECIFIC1 to the sprem cell surface, allowing gamete fusion.

77. Grossniklaus U: **Plant germline development: a tale of cross-talk, signaling, and cellular interactions**. *Sex Plant Reprod* 2011, **24**:91-95.

78. •• Olmedo-Monfil V, Duran-Figueroa N, Arteaga-Vazquez M, Demesa-Arevalo E, Autran D, Grimanelli D, Slotkin RK, Martienssen RA, Vielle-Calzada JP: **Control of female gamete formation by a small RNA pathway in *Arabidopsis***. *Nature* 2010, **464**:628-632.
This report reveals the role of epigenetic factors *AGO9*, *SGS3* and *RDR6* (all involved in small RNA pathways) during female germline specification. Its results suggest that these genes act to restrict the germline to a single cell, such that mutants in these genes result in a phenotype resembling apospory. Furthermore, in *ago9* mutants transposable elements become expressed during megagametogenesis, suggesting that small pathways are involved in the protection of the plant germline against transposition.

79. Li S, Pandey S, Gookin TE, Zhao Z, Wilson L, Assmann SM: **Gene-sharing networks reveal organizing principles of transcriptomes in *Arabidopsis* and other multicellular organisms**. *Plant Cell* 2012, **24**:1362-1378.

80. Chen C, Farmer AD, Langley RJ, Mudge J, Crow JA, May GD, Huntley J, Smith AG, Retzel EF: **Meiosis-specific gene discovery in plants: RNA-Seq applied to isolated *Arabidopsis* male meiocytes**. *BMC Plant Biol* 2010, **10**:280.

81. Yang H, Lu P, Wang Y, Ma H: **The transcriptome landscape of *Arabidopsis* male meiocytes from high-throughput sequencing: the complexity and evolution of the meiotic process**. *Plant J* 2011, **65**:503-516.

82. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**:511-515.

83. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC: **Genome-wide mapping of alternative splicing in *Arabidopsis thaliana***. *Genome Res* 2010, **20**:45-58.

84. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in *Arabidopsis***. *Cell* 2008, **133**:523-536.

85. Song BH, Mitchell-Olds T: **Evolutionary and ecological genomics of non-model plants**. *J Syst Evol* 2011, **49**:17-24.

86. Szövényi P, Rensing SA, Lang D, Wray GA, Shaw AJ: **Generation-biased gene expression in a bryophyte model system**. *Mol Biol Evol* 2011, **28**:803-812.

87. Der JP, Barker MS, Wickett NJ, dePamphilis CW, Wolf PG: ***De novo* characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum***. *BMC Genomics* 2011, **12**:99.

88. Thiel J, Hollmann J, Rutten T, Weber H, Scholz U, Weschke W: **454 transcriptome sequencing suggests a role for two-component signalling in cellularization and differentiation of barley endosperm transfer cells**. *PLoS One* 2012, **7**:e41867.

89. Bargmann BO, Birnbaum KD: **Fluorescence activated cell sorting of plant protoplasts**. *J Vis Exp* 2010, **36**:1673.

8.7 Plant germline formation: molecular insights define common concepts and illustrate developmental flexibility in apomictic and sexual reproduction

The following review is published in "Development" (published by The Company of Biologists Ltd, all rights reserved)[1]. I created all figures and helped to improve parts of the text.

[1]Schmidt, A, Schmid, MW, and Grossniklaus, U (2015) Plant germline formation: molecular insights define common concepts and illustrate developmental flexibility in apomictic and sexual reproduction. Development 142: 229–241.

© 2015. Published by The Company of Biologists Ltd | Development (2015) 142, 229-241 doi:10.1242/dev.102103

REVIEW

Plant germline formation: common concepts and developmental flexibility in sexual and asexual reproduction

Anja Schmidt*, Marc W. Schmid and Ueli Grossniklaus*

ABSTRACT

The life cycle of flowering plants alternates between two heteromorphic generations: a diploid sporophytic generation and a haploid gametophytic generation. During the development of the plant reproductive lineages – the germlines – typically, single sporophytic (somatic) cells in the flower become committed to undergo meiosis. The resulting spores subsequently develop into highly polarized and differentiated haploid gametophytes that harbour the gametes. Recent studies have provided insights into the genetic basis and regulatory programs underlying cell specification and the acquisition of reproductive fate during both sexual reproduction and asexual (apomictic) reproduction. As we review here, these recent advances emphasize the importance of transcriptional, translational and post-transcriptional regulation, and the role of epigenetic regulatory pathways and hormonal activity.

KEY WORDS: Cell fate acquisition, Gene regulation, Germline development, Plant reproduction, Polarity

Introduction

In higher plants, diverse and versatile strategies have evolved to ensure reproductive success. During gametogenesis (see Glossary, Box 1), the male (pollen) and female (embryo sac) gametophytes, which harbour the male (sperm) and female (egg and central cell; see Glossary, Box 1) gametes, respectively, form in specialized reproductive tissues of the flower: the anther and ovule (Fig. 1). The multicellular gametophytes are formed following meiosis of spore mother cells (see Glossary, Box 1), thus producing reduced gametes that harbour half the chromosome number of the maternal sporophyte (haploid in case of diploid plants). During sexual reproduction (see Glossary, Box 1), sperm cells fuse with both the egg and the central cell in the process of double fertilization, giving rise to the embryo and endosperm, respectively, the major components of the seed (Fig. 1). The embryo constitutes the next sporophytic generation, while the endosperm is a terminal nourishing tissue for the embryo and also provides the majority of calories for human and animal consumption. Haploid plants can also form directly from male and female gametes. While this process occurs at very low frequencies in nature, it can be induced in culture and by mutation, and hence is being exploited to accelerate plant breeding (Germanà, 2011). By contrast, during vegetative reproduction (see Glossary, Box 1) and somatic embryogenesis (see Glossary, Box 1), which are two distinct types of asexual reproduction (see Glossary, Box 1), new plants develop without the formation of gametes and seeds. However, plants can also produce seeds via asexual reproduction, avoiding the need for fertilization, in a process known as gametophytic apomixis (hereafter referred to as apomixis, see Glossary, Box 1). Apomixis occurs in more than 400 plant species belonging to ~40 genera.

Both sexual reproduction and apomixis have distinct advantages for natural plant populations and agricultural applications. Sexual reproduction leads to genetically and phenotypically variable offspring, thus forming the basis for plant adaptation to changing environments and allowing for the breeding of new varieties. By contrast, apomixis produces clonal offspring that are genetically identical to the mother plant, thus fixing complex genotypes. Although apomixis is rare among crop plants, the engineering of apomictic crops promises great potential and economical value for crop production and for other applications in agriculture (Koltunow et al., 1995; Vielle-Calzada et al., 1996; Grossniklaus et al., 1998a,b; Spillane et al., 2004).

Over the past decade, plant sexual and apomictic germline formation has attracted the attention of scientists for a number of reasons: (1) the transition from sporophytic to reproductive fate by

Box 1. Glossary

Apomeiosis. The omission or abortion of meiosis during sporogenesis
Apomictic initial cell (AIC). The first cell in the apomictic female germline that omits or aborts meiosis
Apomixis. Asexual reproduction via seed formation
Apospory. The formation of an unreduced female gametophyte from an apomictic initial cell (AIC) developing adjacent to the sexual germline in the ovule
Archesporial cell. The cell giving rise (with or without division) to the spore mother cell
Asexual reproduction. Reproduction without the fusion of gametes
Central cell. The female gamete giving rise to the endosperm
Diplospory. The apomeiotic formation of an unreduced female gametophyte from an AIC at the position of the megaspore mother cell
Egg cell. The female gamete giving rise to the embryo
Functional megaspore (FMS). The cell that develops into the female gametophyte
Gametogenesis. The development of gametophytes from spores
Parthenogenesis. The formation of an embryo from an unfertilized egg cell
Pseudogamy. The fertilization-dependent formation of endosperm from a central cell in apomicts
Sexual reproduction. The mode of reproduction whereby female (egg) and male (sperm) gametes fuse to form a zygote
Somatic embryogenesis. The formation of an embryo from a sporophytic cell without gamete and seed formation
Sporogenesis. The formation of spores from spore mother cells
Spore mother cell. The first cell of the reproductive lineage, formed from sporophytic cells in female and male reproductive tissues of the flower
Synergid cells. Accessory cells of the mature female gametophyte that are important for pollen tube guidance and reception
Vegetative reproduction. A form of reproduction in which a new plant is formed without the formation of an embryo

Institute of Plant Biology and Zürich-Basel Plant Science Centre, University of Zürich, Zollikerstrasse 107, Zürich CH-8008, Switzerland.

*Authors for correspondence (aschmidt@botinst.uzh.ch; grossnik@botinst.uzh.ch)

Development (2015) 142, 229-241 doi:10.1242/dev.102103

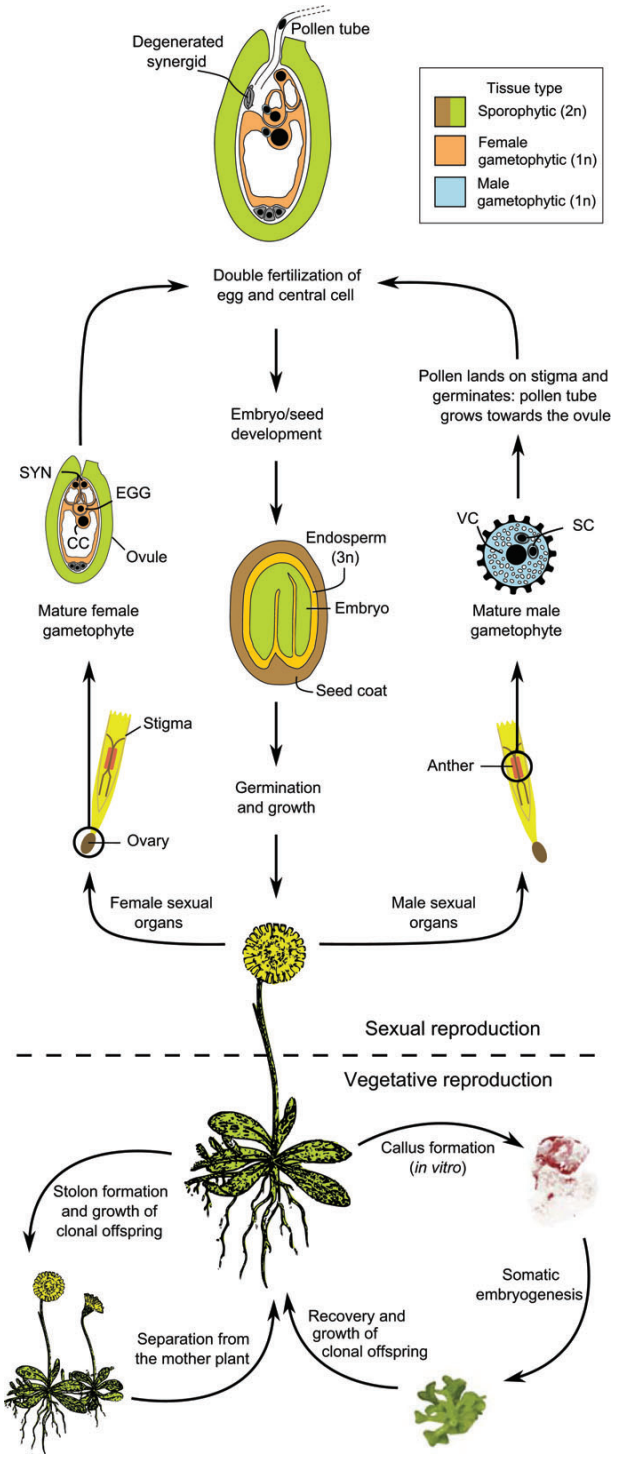


Fig 1. The life cycle of a plant. Plants have a more complex life cycle than animals, alternating between two heteromorphic generations: the sporophyte and the gametophyte. In the diploid sporophyte, distinct cells undergo meiosis and produce haploid spores. These give rise to multicellular haploid gametophytes, which produce gametes through mitotic divisions. The fusion of a male (sperm) and a female (egg) gamete results in the formation of a zygote, which constitutes the sporophyte. The example depicted, *Hieracium pilosella*, follows the common life cycle of angiosperms. The anthers of the flower produce the male gametophyte (called the pollen or microgametophyte), which consists of three cells: two sperm cells (SC) and one vegetative cell (VC). The female gametophyte (called the embryo sac or megagametophyte) is embedded in maternal sporophytic tissues of the ovule. The latter is enclosed in the carpels of the flower. For sexual reproduction, the pollen needs to germinate on the stigma and to deliver the two sperm cells to the female gametophyte. Fertilization, the transition from the gametophytic to the sporophytic generation, occurs within the ovule. In addition to sexual reproduction, plants can frequently reproduce asexually, e.g. by stolon outgrowth (vegetative reproduction) or via the formation of calli in culture, followed by somatic embryogenesis and development into an adult plant. CC, central cell; EGG, egg cell; SC, sperm cell; SYN, synergid cell; VC, vegetative cell.

reprogramming a somatic cell is a key step in the plant life cycle; (2) during gametogenesis, a few rounds of mitosis and cellularization lead to the formation of functionally distinct cell types that are all derived from a single spore, a process ideally suited to address fundamental questions in developmental biology; and (3) understanding the molecular mechanisms that determine sexual or asexual fate decisions is a precondition for the targeted manipulation of plant reproduction for agricultural use and crop improvement. Accordingly, many studies have focussed on determining the gene expression profiles, epigenetic mechanisms and regulatory pathways involved in germline development (reviewed by Drews and Koltunow, 2011; Sprunck and Gross-Hardt, 2011; Schmidt et al., 2012; Gutierrez-Marcos and Dickinson, 2012; Wüest et al., 2013). Here, we focus on recent studies that have elucidated the molecular mechanisms underlying the acquisition of reproductive fate in sexual and apomictic species, the determination of meiosis versus apomeiosis (see Glossary, Box 1), and the polar development of the female gametophyte.

Development of the plant reproductive lineages

The formation of the plant reproductive lineages proceeds in two distinct phases: during sporogenesis (Glossary, Box 1), spores are formed by sporophytic (somatic) cells, whereas during gametogenesis the spores develop into mature gametophytes that harbour the male or female gametes (Fig. 2). During the course of evolution, the gametophytic phase of the plant life cycle, which is dominant in bryophytes (i.e. liverworts, hornworts and mosses), has been dramatically reduced to only a few cells in the angiosperms (flowering plants). Thus, unlike in most animals, where the germline is set aside early in embryogenesis, the plant germline is determined only late in development, during floral organ formation. Here, we consider the spore mother cells to be the first cells of the germline, as the lineage of the gametes can unambiguously be traced back to them (Grossniklaus, 2011). However, it should be noted that, because gametophytes consist of both gametic and non-gametic accessory cells and the germline is defined as the cell lineage that differentiates into gametes, some authors place the determination of the germline later during gametophyte development to the immediate precursors of the gametes (e.g. Berger and Twell, 2011; Twell, 2011).

The formation of the male reproductive lineage begins with the differentiation of a microspore mother cell (MiMC) in the developing anthers; the periclinal division of archesporial cells (see Glossary, Box 1) gives rise to outer parietal cells and inner sporogenous cells, and the MiMCs differentiate from the latter. The MiMC undergoes meiosis to give rise to a tetrad of microspores (Fig. 2), each of which undergoes an asymmetric division (termed pollen mitosis I, PMI) to form a vegetative and a generative cell (Borg et al., 2009). During pollen mitosis II (PMII), the generative cell forms two sperm cells (male gametes), while the vegetative cell does not divide again. The

Development (2015) 142, 229-241 doi:10.1242/dev.102103

sperm cells are then delivered to the female gametes by the pollen tube, which forms via growth of the vegetative cell. The timing of PMII varies in different species; in most plant species, PMII takes place in the growing pollen tube but in some species, including *Arabidopsis* and maize, the generative cell divides before the pollen is released from the anther (Boavida et al., 2005).

During formation of the female sexual reproductive lineage, typically a single somatic cell per ovule acquires reproductive fate and differentiates to form an archesporial cell. It can be distinguished from the surrounding cells by its subepidermal localization and its enlarged size. In the sexual model species *Arabidopsis*, as in most species, the archesporial cell directly differentiates into a megaspore mother cell (MMC) without intervening divisions. The MMC is defined by its commitment to the meiotic fate and gives rise to a tetrad of megaspores (Fig. 2). Typically, only one functional megaspore (FMS; Glossary, Box 1) survives while the others degenerate (Fig. 2). Interestingly, the FMS occupies a defined position in the ovule, suggesting that this is important for its survival and cell fate acquisition. A role for signalling from sporophytic ovule tissues during the selection of the FMS has been discussed (Koltunow, 1993; Grossniklaus and Schneitz, 1998; Koltunow and Grossniklaus, 2003) and, in maize, the accumulation of callose in the cell walls of the degenerating megaspores has been hypothesized to play a role in shielding these cells from such signals (Russell, 1979). The FMS, in turn, typically undergoes three mitotic divisions to form a syncytial female gametophyte (Fig. 2). In most species, cellularization results in an eight-nucleate, seven-celled mature gametophyte (embryo sac), referred to as a *Polygonum* type embryo sac. It harbours the two female gametes, the synergid cells (see Glossary, Box 1), which are important for pollen tube guidance and reception, and three antipodal cells (Fig. 2). Although the role of the antipodal cells remains unclear, they might be involved in transferring nutrients from the surrounding sporophytic tissues to the embryo sac (Raghavan, 1997). The *Polygonum* type embryo sac occurs in ~70% of all angiosperms, including the model systems *Arabidopsis thaliana* (mouse ear cress), *Zea mays* (maize) and *Oryza sativa* (rice), and many apomictic species (Drews and Koltunow, 2011).

From the beginning of its development, the female gametophyte is highly polarized, suggesting that positional information may play a role in cell fate acquisition (Grossniklaus and Schneitz, 1998; Lituiev and Grossniklaus, 2014). The exact position of nuclei within the syncytium may thus be an important factor for cell specification during cellularization (Sundaresan and Alandete-Saez, 2010; Sprunck and Gross-Hardt, 2011). It is also evident that variations in this developmental pattern exist: while megasporogenesis typically leads to a single surviving one-nucleate FMS (monosporic megasporogenesis), failures in cell plate formation after meiosis I or after both meiotic divisions can lead to two- or four-nucleate FMSs, developmental patterns referred to as bisporic or tetrasporic megasporogenesis, respectively (Maheshwari, 1950; Willemse and Went, 1984; Haig, 1990; Huang and Russell, 1992; Drews and Koltunow, 2011). Other developmental variations concern the number of mitoses during megagametogenesis before cellularization, the possibility of additional mitoses after cellularization, and the timing of the fusion of the polar nuclei in the central cell (Maheshwari, 1950; Drews and Koltunow, 2011).

Sexual reproduction and apomixis are interrelated

Compared with sexual reproduction, apomixis differs only in three key developmental steps (Fig. 2). First, female meiosis is circumvented, in a process referred to as apomeiosis, leading to the formation of unreduced megaspores and, consequently, unreduced female gametes. The first cell of the apomictic lineage is termed an apomictic initial cell (AIC; see Glossary, Box 1). The AIC is either formed at the position of the MMC and omits or aborts meiosis (diplospory; see Glossary, Box 1) to give rise to an unreduced FMS, or is derived from a somatic cell in close proximity to the MMC that directly differentiates into an unreduced FMS (apospory; see Glossary, Box 1) (Bicknell and Koltunow, 2004). Usually, male meiosis is unaffected but unreduced pollen can also be produced in some apomicts (Bicknell and Koltunow, 2004). Second, the egg develops into an embryo in the absence of fertilization in a process known as parthenogenesis (see Glossary, Box 1). Currently, the molecular mechanisms that activate the egg cell and initiate embryogenesis are unknown. Third, the central cell can form endosperm either autonomously or after fertilization (pseudogamy; see Glossary, Box 1). Functional endosperm formation in pseudogamous apomicts requires adaptations in either megagametogenesis (the production of a four-nucleate embryo sac), microgametogenesis (the formation of unreduced sperm cells) or double fertilization to ensure a balanced endosperm with the correct 2:1 ratio of maternal to paternal genomes crucial for seed development in many species (Grossniklaus, 2001; Koltunow and Grossniklaus, 2003; Spillane et al., 2004). For example, in maize *indeterminant gametophyte1* (*ig1*) mutants, abnormal numbers of nuclei are formed in the female gametophyte, leading to an aberrant maternal to paternal genome ratio in the endosperm, which results in seed abortion (Lin, 1984; Huang and Sheridan, 1996). In autonomous apomicts, the requirement for a balanced endosperm is alleviated, likely also depending on specific adaptations that are under genetic control.

The acquisition and restriction of reproductive fate

During sexual reproduction, only one somatic cell per ovule is usually committed to the reproductive fate. However, it is not fully understood what determines the commitment of this somatic cell to initiate germline development and what prevents the formation of additional germline cells in the same ovule. In higher plants, it has been hypothesized that the MMC represses the formation of additional MMCs and thus restricts germline formation to only one cell per ovule (Grossniklaus and Schneitz, 1998). In support of the hypothesis that the germline itself suppresses the formation of additional germline lineages, the formation of multiple female gametophytes per ovule has been described in *Trimenia*, an ancient angiosperm taxon, where tip growing female gametophytes compete to reach the site of fertilization (Bachelier and Friedman, 2011).

Initial insights into the signalling pathways that regulate the restriction of germline fate came from the analyses of mutants in maize, rice and *Arabidopsis* (summarised in Table 1). In rice carrying mutations in *MULTIPLE SPOROCYTE* (*MSP1*) and in *Arabidopsis* plants carrying mutations in the orthologue *EXTRA SPOROGENOUS CELLS*/*EXCESS MICROSPOROCYTES1* (*EXS*/*EMS1*) or mutations in *SOMATIC EMBRYOGENESIS RECEPTOR KINASE1* and *2* (*SERK1*/*2*), more MiMCs develop per anther in comparison to the wild type (Canales et al., 2002; Zhao et al., 2002; Nonomura et al., 2003; Albrecht et al., 2005; Colcombet et al., 2005; Jia et al., 2008). These genes encode leucine-rich receptor kinases (*MSP1* and *EXS*/*EMS1*) and LRR receptor-like serine threonine kinases (*SERK1*/*2*) (Canales et al., 2002; Zhao et al., 2002; Nonomura et al., 2003; Albrecht et al., 2005; Colcombet et al., 2005; Jia et al., 2008). Similar phenotypes have been described in mutant in which the genes *TAPETUM DETERMINANT1* (*TPD1*) in

Development (2015) 142, 229-241 doi:10.1242/dev.102103

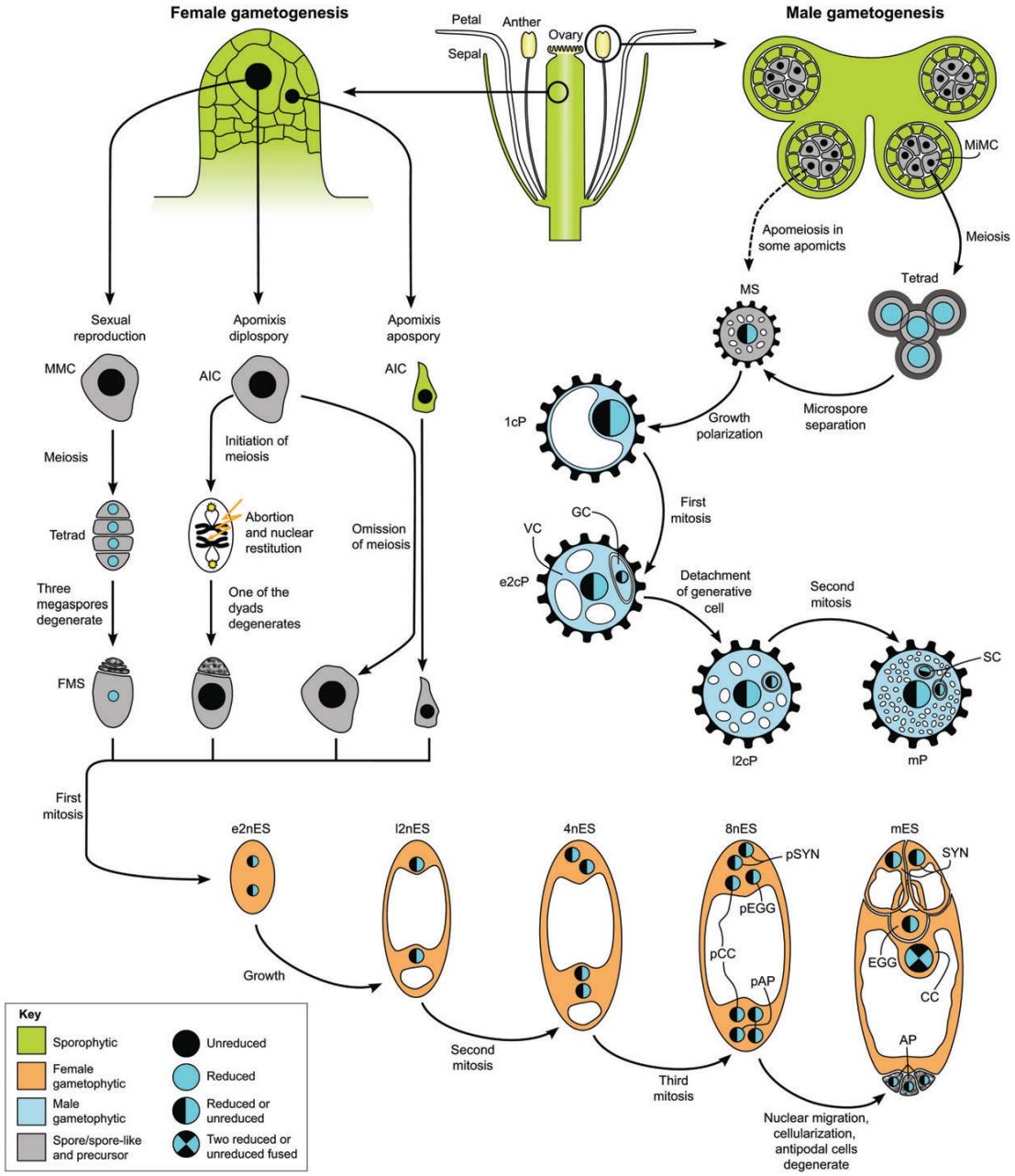


Fig. 2. Male and female gametophyte development in sexually reproducing and apomictic plant species. Germline development starts with the differentiation of sporophytic cells into spore mother cells (female, megaspore mother cell, MMC; male, microspore mother cell, MiMC) that, in sexually reproducing species undergo meiosis to give rise to four haploid spores. During male gametogenesis, the four spores separate and form unicellular microspores (MS), which grow by cell expansion to form unicellular pollen (1cP). The first asymmetric mitosis produces bicellular pollen (e2cP) containing a large vegetative cell (VC) and a small generative cell (GC). The GC detaches from the cell wall and becomes engulfed by the VC. Sperm cells (SC) are formed during the second mitosis of the GC. The mature pollen (mP) consists of a VC, which will form the pollen tube, and two SCs that mediate double fertilization. During female gametogenesis, three of the four spores degenerate, leaving one functional megaspore (FMS), which undergoes three mitotic divisions in a syncytium to give rise to the early/late two-nucleate, four-nucleate and then eight-nucleate embryo sac (e/l2nES, 4nES, 8nES). Nuclear migration and concomitant cellularization eventually lead to the formation of a mature embryo sac (mES), a highly polarized structure that contains four distinct cell types: two synergid cells (SYN), the egg cell (EGG), the central cell (CC) and antipodal cells (APs), which degenerate prior to fertilization. In apomictic species, different mechanisms can lead to the formation of unreduced gametes. In diplospory, the apomictic initial cell (AIC) may initiate meiosis but restitution results in the formation of two unreduced AICs, one of which degenerates. By contrast, aposporous apomicts form a FMS-like cell at a different position in the ovule. The unreduced AICs then develop into unreduced female gametophytes. Meiosis on the male side is usually normal in apomicts. Female gametophyte stages (FG) are according to Christensen et al. (1997). p, precursor of.

Development (2015) 142, 229-241 doi:10.1242/dev.102103

Arabidopsis, *MULTIPLE ARCHESPORIAL CELLS* (*MAC1*) in maize and its rice orthologue *OsTDL1A* are disrupted. These genes encode small secreted proteins identified as the putative ligands of the MSP1 or EXS/EMS1 receptor kinases (Sheridan et al., 1996, 1999; Yang et al., 2003, 2005; Zhao et al., 2008; Wang et al., 2012; Kelliher and Walbot, 2012). Unlike in *Arabidopsis*, in rice and maize this pathway also affects female sporogenesis, indicating differences in the mechanism of repression of additional sporocytes (Sheridan et al., 1996; Zhao et al., 2008; Nonomura et al., 2003). In maize and rice, excess archesporial cells were observed, leading to the formation of more sporogenous cells (Zhang and Yang, 2014). In *Arabidopsis*, however, the pathway plays a role in cell fate decisions and cell specification after the periclinal division of the archesporial cell. As recently demonstrated, EXS/EMS1 forms complexes with SERK1/2 to control the proliferation of tapetal cells in the anther (Albrecht et al., 2005; Colcombet et al., 2005; Feng and Dickinson, 2010). Interestingly, partially complementary expression patterns have been reported for *TPD1* and *EXS*/*EMS1*, which are predominantly expressed in sporogenous cells and tapetal cells, respectively, at the developmental stages at which the mutant phenotypes are first established, indicating signalling between cell types (Yang et al., 2003).

In aposporous apomicts, such repression of additional germline lineages is not active, as both an enlarged somatic AIC and the MMC can initiate reproductive lineages (Fig. 1). Formation of the AIC in *Hieracium pilosella* even depends on differentiation and meiosis of the sexual MMC (Koltunow et al., 2011). Thus, as suggested by the signalling pathways described above, communication between cell types during sporogenesis seems to be involved in cell type specification and the acquisition or restriction of germline fate. It remains unclear whether this is achieved by overcoming the mechanism that usually represses the development of additional germline cells or by an alternative signalling pathway that induces reproductive fate in an additional somatic cell, or whether a combination of both of these mechanisms is involved. However, it should be noted that, once established, the apomictic lineage often suppresses the further development of the sexual female gametophyte (Koltunow et al., 2011), suggesting that distinct control mechanisms exist at these developmental steps.

In diplosporous species, the AIC omits or aborts meiosis producing an unreduced FMS (Fig. 2; Bicknell and Koltunow, 2004). As in sexual species, diplosporous apomicts typically develop only one germline lineage per ovule, suggesting that the processes by which gametophytic fate is acquired in apospory and diplospory follow distinct developmental principles. It is unknown whether apomictic fate is regulated by related or alternative molecular mechanisms in diplosporous and aposporous species. Investigations into this question have proved to be technically challenging as the female germline is deeply embedded in maternal floral tissues. Nevertheless, recent methodological advances have allowed the transcriptional profiling of such rare cell types by combining laser-assisted microdissection or micromanipulation with microarray and/or RNA-Seq analyses. These studies have provided novel insights into the transcriptional basis of germline specification and development (Wüest et al., 2010; Schmidt et al., 2011, 2012, 2014; Schmid et al., 2012; Okada et al., 2013; Wüest et al., 2013; Abiko et al., 2013; Anderson et al., 2013; Chettoor et al., 2014).

New insights into the issue of how cell specification is regulated during diplospory, when compared with sexual or aposporous reproduction, were recently provided by cell type-specific transcriptome analyses of the reproductive lineage in *Boechera gunnisoniana*, a diplosporous apomict that is related to sexual *A. thaliana* (Schmidt et al., 2014). Comparative transcriptome analyses detected a number of commonalities between the sexual MMC and the diplosporous AIC (Schmidt et al., 2014). Importantly, significant differences in the activities of a number of regulatory pathways were also observed, including differences in cell cycle regulation, hormonal pathways, signal transduction and epigenetic regulatory pathways (Schmidt et al., 2014). Through comparisons with a transcriptome dataset of the AIC of *Hieracium praealtum* (Okada et al., 2013), this study suggests interesting differences between the regulatory mechanisms specifying a diplosporous or an aposporous AIC (Schmidt et al., 2014). Importantly, the *H. praealtum* AIC seems to have already adopted a gametophytic fate (Okada et al., 2013). In agreement with this acquisition of a FMS fate without meiotic division, a number of meiotic genes are not expressed in the *H. praealtum* AIC (Okada et al., 2013). By contrast, the majority of 25 core meiotic genes are expressed in the AIC of

Table 1. Genes involved in the restriction of additional sporogenous cells during plant reproductive development

Gene	Species	Restricts number of MiMCs, MMCs or both	Type/function of protein encoded	Reference(s)
MSP1	*O. sativa*	Restricts the number of sporocytes in anther and ovule	Leucine-rich repeat receptor-like kinase; orthologue of *Arabidopsis EXS*/*EMS1*	Nonomura et al., 2003
EXS/*EMS1*	*A. thaliana*	Restricts the number of microsporocytes	Leucine-rich repeat receptor-like kinase involved in regulating the proliferation of tapetal cells during anther development	Canales et al., 2002; Zhao et al., 2002; Feng and Dickinson, 2010
SERK1/2	*A. thaliana*	Restricts the number of microsporocytes	Leucine-rich repeat receptor-like kinases; forms complexes with EXS/EMS1 in tapetal cells	Albrecht et al., 2005; Colcombet et al., 2005
TPD1	*A. thaliana*	Restricts the number of microsporophytes	Small secreted protein; interacts with EXS/EMS1	Yang et al., 2003; Yang et al., 2005
MAC1	*Z. mays*	Restricts the number of archesporial cells in anther and ovule	Small secreted protein; orthologue of Os*TDL1A*	Sheridan et al., 1996; Sheridan et al., 1999; Wang et al., 2012
Os*TDL1A*	*O. sativa*	Restricts the number of sporocytes in anther and ovule	Small secreted protein; putative ligand of MSP1	Zhao et al., 2008

A. thaliana, *Arabidopsis thaliana*; *EXS*/*EMS1*, *EXTRA SPOROGENOUS CELLS*/*EXCESS MICROSPOROCYTES1*; *MAC1*, *MULTIPLE ARCHESPORIAL CELLS1*; MiMCs, microspore mother cells; MMCs, megaspore mother cells; *MSP1*, *MULTIPLE SPOROCYTE1*; *O. sativa*, *Oryza sativa*; Os*TDL1A*, orthologue of *MAC1*; *SERK1/2*, *SOMATIC EMBRYOGENESIS RECEPTOR KINASE1* and *2*; *TPD1*, *TAPETUM DETERMINANT1*; *Z. mays*, *Zea mays*.

the diplosporous species *B. gunnisoniana* before first division restitution (Schmidt et al., 2014). This supports the notion that diplospory results from a modification of the meiotic pathway in an MMC-like cell, while the aposporous AIC becomes directly determined to a gametophytic fate without prior activation of the meiotic program.

Mutations in meiotic genes can lead to diplospory-like modifications of meiosis

Over recent years, investigations into the regulatory processes governing meiosis have allowed the identification of meiotic mutants that generate unreduced gametes (Table 2) (Brownfield and Köhler, 2011; Crismani et al., 2013). For example, mutations in the gene encoding *DYAD*/*SWITCH1* (*SWI1*) lead to apomeiosis and to the formation of rare triploid offspring that retain full parental heterozygosity (Ravi et al., 2008). In *MiMe-1* and *MiMe-2* triple mutants, a diplospory-like division also leads to the formation of unreduced gametes. *MiMe-1* and *MiMe-2* are combinations of *sporulation11-1* (*spo11-1*), *omission of second division1* (*osd1*) and *recombination8* (*rec8*), and *spo11-1*, *osd1* and *cyc1;2*/*tardy asynchronous meiosis* (*tam*), respectively (d'Erfurth et al., 2009, 2010). Using these *Arabidopsis* mutants, synthetic clonal seeds have been produced by manipulating the expression of the centromere-specific histone 3 variant CENH3, which leads to paternal genome elimination, in the *dyad/swi1* or *MiMe* mutant background (Marimuthu et al., 2011).

Meiosis and the acquisition of germline fate are affected by abiotic and oxidative stress

Although these mutations in meiotic genes lead to a deregulation of the meiotic program and, eventually, to a switch to a diplospory-like process, little is known about the control of diplospory and meiotic restitution in natural apomicts. Interestingly, abiotic stress can lead to alterations in meiotic cell division (de Storme and Geelen, 2014). For example, in rose (*Rosa spp.*) short periods of heat stress result in partial restitution of male meiosis and the formation of unreduced dyads, but also triads and polyads (Pecrix et al., 2011).

Analysis of the effect of redox status on germline specification in maize revealed another link to abiotic stress. In anthers, germ cell formation is stimulated by a low oxygen environment or by a low abundance of reactive oxygen species (ROS), which accumulate under different kinds of stress (Kelliher and Walbot, 2012, 2014). This led to the conclusion that reduced oxygen concentration promotes the acquisition of meiotic fate in maize (Kelliher and Walbot, 2012). Contrasting the idea that meiotic fate is acquired under low ROS levels, a recent hypothesis postulates that the evolution and maintenance of meiosis depends on stress and elevated ROS levels (Hörandl and Hadacek, 2013, see Box 2).

In conclusion, although abiotic and oxidative stresses seems to play a role in the transition from somatic to reproductive fate and the regulation of (apo)meiosis, their potential role as a driving force promoting sexual or asexual reproduction remains unclear and warrants further investigation.

Epigenetic regulatory pathways are important for germline specification and the control of sexual versus apomictic reproduction

Disturbance of the meiotic programme typically results in sterility or the diplospory-like formation of unreduced gametes. However, phenotypes resembling apospory or diplospory have also been observed in mutants perturbing epigenetic regulatory pathways, in particular those involving DNA methylation and small RNA-based gene regulation (Olmedo-Monfil et al., 2010; Garcia-Aguilar et al., 2010; Singh et al., 2011).

Epigenetic regulation is involved in a variety of developmental and cell fate decisions by controlling gene activity through DNA or chromatin modifications. For example, ARGONAUTE (AGO) proteins are involved in gene regulation mediated by small RNAs such as microRNAs (miRNA), small interfering RNAs (siRNA) and PIWI-associated RNAs (piRNA) (Meister, 2013). In *Arabidopsis*, 10 AGO proteins have been identified, and these can be grouped into three major clades: the AGO1, AGO5 and AGO10 clade; the AGO2, AGO3 and AGO7 clade; and the AGO4, AGO6, AGO8 and AGO9 clade (Mallory and Vaucheret, 2010). These different clades of AGO proteins engage in different small RNA pathways, with the AGO9 clade being active in the siRNA heterochromatin pathway that regulates the transcriptional silencing of transposons and repeats by mediating DNA methylation and heterochromatin formation (Mallory and Vaucheret, 2010).

Table 2. Mutations that lead to the formation of unreduced female gametophytes by an apospory- or diplospory-like mechanism

Mutation	Species	Description	Type of apomeiosis	Reference(s)
dyad/*swi1*	*A. thaliana*	Mutation in core meiotic gene	Diplospory like	Ravi et al., 2008
MiMe-1 (*spo11-1*, *osd1* and *rec8*)	*A. thaliana*	Triple mutant of core meiotic genes	Diplospory like	d'Erfurth et al., 2009
MiMe-2 (*spo11-1*, *osd1* and *tam*)	*A. thaliana*	Triple mutant of core meiotic genes	Diplospory like	d'Erfurth et al., 2010
ago9	*A. thaliana*	Mutation in gene involved in a small RNA pathway	Apospory like	Olmedo-Monfil et al., 2010
rdr6	*A. thaliana*	Mutation in gene involved in a small RNA pathway	Apospory like	Olmedo-Monfil et al., 2010
sgs3	*A. thaliana*	Mutation in gene involved in a small RNA pathway	Apospory like	Olmedo-Monfil et al., 2010
mem	*A. thaliana*	Mutation in gene encoding a RNA-helicase	Apospory like	Schmidt et al., 2011
dmt102	*Z. mays*	Mutation in gene involved in DNA methylation	Apospory like	Garcia-Aguilar et al., 2010
dmt103	*Z. mays*	Mutation in gene involved in DNA methylation	Apospory like	Garcia-Aguilar et al., 2010
ago104	*Z. mays*	Mutation in gene involved in a small RNA pathway	Diplospory like	Singh et al., 2011

A. thaliana, *Arabidopsis thaliana*; *Z. mays*, *Zea mays*.

Development (2015) 142, 229-241 doi:10.1242/dev.102103

Box 2. The evolution of apomixis and meiosis

Evidence suggest that apomixis evolved from a deregulation of the sexual pathway several times independently (Koltunow, 1993; Vielle-Calzada et al., 1996; Leblanc et al., 1997; Grimanelli et al., 2001; Grossniklaus, 2001; Tucker et al., 2003; Koltunow and Grossniklaus, 2003; Sharbel et al., 2009, 2010). Deregulation of genetic and epigenetic regulatory pathways has been hypothesized to be a consequence of hybridization and polyploidization, which have been proposed as preconditions for apomixis to occur (Asker and Jerling, 1992; Grossniklaus, 2001; Spillane et al., 2001; de Storme and Geelen, 2013).

Interestingly, according to a recent hypothesis, meiosis as a precondition for sexual reproduction is thought to have evolved as a repair mechanism for DNA damage induced by oxidative stress and ROS, and it has been proposed that the redox chemistry between oxidized DNA and the meiotic protein SPO11 is required for the generation of double-strand breaks, which are required for meiotic recombination and the repair of damaged DNA (Hörandl and Hadacek, 2013). However, low levels of ROS promote the acquisition of meiotic fate in maize anthers (Kelliher and Walbot, 2012). In line with this hypothesis, the metabolism of polyamine and spermidine, which are quenchers of ROS activity, is enriched in the AIC in *B. gunnisoniana* (Schmidt et al., 2014). Similarly, increasing evidence suggests the importance of the redox state for the development of the anther and male germline (reviewed by Zhang and Yang, 2014). Nevertheless, the role of stress and reactive oxygen species on regulating sexual versus apomictic reproduction remains unknown.

Increasing evidence highlights the importance of *AGO* activity in plant germline development and gamete formation (Nonomura et al., 2007; Wüest et al., 2010; Olmedo-Monfil et al., 2010; Singh et al., 2011; Borges et al., 2011; Tucker et al., 2012). This is reminiscent of the role of AGO proteins in the animal germline; proteins of the animal-specific PIWI-clade protect the genomic integrity of the germline, in particular by repressing the activity of transposons in invertebrates, although their role in vertebrates is less clear (Clark and Lau, 2014). AGO/PIWI proteins, and potentially other proteins involved in small RNA pathways, interact with VASA or VASA-like RNA helicases that are preferentially expressed in the germline (Yajima and Wessel, 2011). Although neither the PIWI clade of AGOs nor VASA RNA helicases have been identified in plants, recent evidence suggests that similar regulatory mechanisms evolved in the plant reproductive lineage, likely by convergence (Wüest et al., 2010; Schmidt et al., 2011).

In *Arabidopsis*, a role for *AGO9* and *MNEME* (*MEM*), a RNA helicase that is preferentially expressed in the MMC, during germline specification has recently been described (Fig. 3; Table 2; Olmedo-Monfil et al., 2010; Schmidt et al., 2011). In plants heterozygous for *mem-1* or *mem-2*, more than one subepidermal enlarged cell instead of the single MMC develops in ~21% of ovules, whereas 37-48% of the ovules in *ago9* homozygotes show a similar phenotype, depending on the allele (Olmedo-Monfil et al., 2010; Schmidt et al., 2011). In *mem* and *ago9* mutants, these additional subepidermal enlarged cells directly give rise to a female gametophyte in the absence meiosis, closely resembling apospory (Olmedo-Monfil et al., 2010; Schmidt et al., 2011). Although the AGO9 protein has been detected only in the L1 layer of the developing ovule, and not in the MMC, *MEM* transcripts are highly enriched in the MMC but are also detected in the surrounding ovule tissue, albeit at much lower levels (Fig. 3; Olmedo-Monfil et al., 2010; Schmidt et al., 2011). Thus, MEM and AGO9 likely repress the acquisition of reproductive fate in the surrounding tissues in a non-cell-autonomous manner (Olmedo-Monfil et al., 2010; Schmidt et al., 2011). Interestingly, *AGO9* plays a role in repressing transposons in the germline, reminiscent of the role played by PIWI proteins in animals (Olmedo-Monfil et al., 2010). It remains to be determined whether MEM – like VASA in the animal germline – is involved in this process, potentially acting by aiding the unwinding of RNAs prior to their association with AGO proteins.

Similar phenotypes have also been reported for mutations that disrupt *RNA-DEPENDENT RNA POLYMERASE6* (*RDR6*) and *SUPPRESSOR OF GENE SILENCING3* (*SGS3*), which are known to be required for the biogenesis of *trans*-acting siRNAs (Table 2; Olmedo-Monfil et al., 2010). In addition, features of apospory have been reported for the maize *dmt102* and *dmt103* lines, which carry mutations in the homologues of the *Arabidopsis CHROMOMETHYLTRANSFERASE3* (*CMT3*) and *DOMAINS REARRANGED METHYLASE1* (*DRM1*) and *DRM2*, respectively (Table 2; Garcia-Aguilar et al., 2010). Currently, no evidence has been reported that demonstrates the formation of viable offspring from the unreduced, supernumerous gametophytes seen in the maize *dmt102* and *dmt103* mutants, or in the *Arabidopsis ago9* and *mem* mutants. By contrast, a mutation in maize *AGO104*, a homologue of *Arabidopsis AGO9*, leads to features of diplospory and to the formation of tripoid and tetraploid offspring following the fertilization of unreduced gametophytes (Table 2; Singh et al., 2011). The *ago104* phenotype is caused by a mutation that leads to defects in chromosome condensation during meiosis (affecting mega- and microsporogenesis) and, subsequently, to the formation

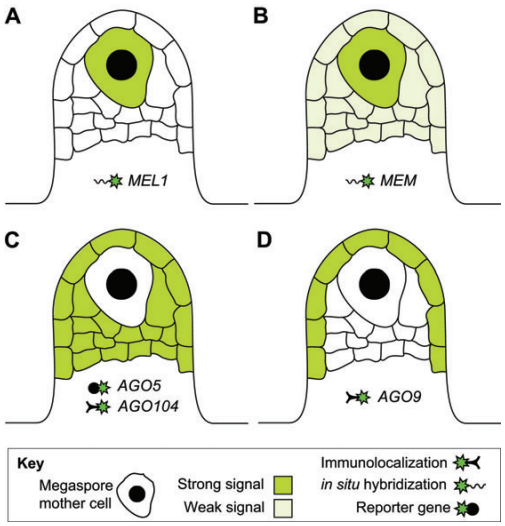


Fig. 3. The expression patterns of proteins/genes whose perturbations mimic apomeiosis. The expression patterns or abundance of protein are schematically shown for: (A) *MEL1* (Nonomura et al., 2007); (B) *MEM* (Schmidt et al., 2011); (C) *AGO5* (Tucker et al., 2012) and AGO104 (Singh et al., 2011); and (D) AGO9 (Olmedo-Monfil et al., 2010). During female germline formation, *MEL1* is expressed in the MMC, suggesting a cell-autonomous effect to cause failure of meiosis (Nonomura et al., 2007). However, a more complex regulation cannot be excluded, as in rare cases germline formation fails. AGO104 and AGO5 are both localized in the nucellus tissue, suggesting signalling from the sporophytic tissues of the nucellus to the developing germline (Singh et al., 2011; Tucker et al., 2012). Interestingly, AGO9 was described to be restricted to the L1 layer of the nucellus, suggesting a non-cell-autonomous mechanism (Olmedo-Monfil et al., 2010). By contrast, highest expression of *MEM* has been observed in the MMC, so that a non-cell-autonomous mechanism to repress germline fate in the surrounding cells is likely. However, as *MEM* is also expressed in the nucellus surrounding the MMC at low levels, other mechanisms cannot be excluded.

of unreduced dyads (Singh et al., 2011). Interestingly, *AGO104* is expressed not in the MMC but in the surrounding somatic tissues, suggesting that the meiotic defect is mediated by a mobile signal (Fig. 3; Singh et al., 2011). Effects of AGO activity on meiosis have been described previously: in rice, mutations in the MMC-expressed gene *MEIOSIS ARRESTED AT LEPTOTENE1* (*MEL1*) (Fig. 3) lead to meiotic arrest and sterility (affecting mega- and microsporogenesis) (Nonomura et al., 2007). MEL1 is closely related only to *Arabidopsis* AGO5 and thus belongs to a different AGO clade than AGO104 (Nonomura et al., 2007). Together, the data suggest diverse and important functions for AGO proteins that are active in different small RNA-dependent regulatory pathways during germline specification and meiosis in different plant species.

Predominant expression of *AGO1*, *AGO2*, *AGO5*, *AGO8* and *AGO9* has also been observed in the *Arabidopsis* egg cell (Wüest et al., 2010), although the function of these AGO proteins in the egg cell remains to be elucidated. *AGO5* is also highly enriched in sperm (Borges et al., 2011), and a role for AGO5 in a putative miRNA complex in the male germline has been proposed (Borges et al., 2011). During megasporogenesis, *AGO5* can be detected in sporophytic ovule tissues, but not in the developing female germline, similar to *AGO9* (Fig. 3; Tucker et al., 2012). Furthermore, plants carrying the semi-dominant *ago5-4* allele do not initiate female gametophyte development, suggesting that this particular mutation inhibits a somatic small RNA pathway that promotes the initiation of gametogenesis (Tucker et al., 2012).

In addition to involving DNA methylation and small RNA pathways, epigenetic pathways regulate chromatin organization and histone modifications. It was recently reported that large-scale chromatin reprogramming establishes an epigenetic and transcriptional state in the *Arabidopsis* MMC that is distinct from that in the surrounding tissue (She et al., 2013). These changes likely contribute to the acquisition of germline fate and to the transition to the gametophytic phase, rather than being only a precondition for meiosis (She et al., 2013). This is supported by the finding of similar histone modifications and histone variant dynamics in the additional subepidermal enlarged cells in *ago9*, *sds3* and *rdr6* mutants (She et al., 2013).

In conclusion, epigenetic regulatory pathways play important roles during the acquisition of germline fate, during germline differentiation and for discriminating between a meiotic and a mitotic fate. It remains unknown whether the influence of stress on germline specification described above acts via changes in the activity of epigenetic pathways or through independent mechanisms.

Polarity and cell fate determination during megagametogenesis

Whether generated by sexual reproduction or apomixis, a highly polarized structure harbouring functionally distinct cell types is established during megasporogenesis from a single FMS by only two or three mitotic divisions and cellularization. Recent studies in *Arabidopsis* have identified a number of factors that can influence this polarity and the subsequent development of the gametophyte (summarized in Table 3; reviewed by Sundaresan and Alandete-Saez, 2010; Sprunck and Gross-Hardt, 2011; Lituiev and Grossniklaus, 2014).

Factors regulating FMS selection

To initiate megagametogenesis, typically only the chalazal-most spore in the tetrad survives and differentiates into the FMS, but the mechanism governing FMS selection and survival is unclear. The *Arabidopsis antikevorkian* mutant affects FMS selection but the corresponding gene remains to be cloned (Yang and Sundaresan, 2000). More recently, *ARABINOGALACTAN PROTEIN18* (*AGP18*) was found to be important for the survival and selection of the FMS (Table 3; Acosta-García and Vielle-Calzada, 2004; Demesa-Arévalo and Vielle-Calzada, 2013). Overexpression of *AGP18* in ovules results in the survival of more than one of the four megaspores (Demesa-Arévalo and Vielle-Calzada, 2013). The mechanism by which *AGP18* determines FMS selection remains unknown, although it has been hypothesized that AGP proteins, which are attached to the plasma membrane through a glycosylphosphotidylinositol (GPI) anchor, can act as components of signalling pathways (Youl et al., 1998; Borner et al., 2003; Ellis et al., 2010; Seifert and Roberts, 2007; Zhang et al., 2011).

Subsequent to megaspore selection, the FMS forms the mature gametophyte, which harbours four functionally distinct cell types, typically through three mitotic divisions. How cell fate acquisition is regulated and when cell fate is determined during this process is still largely unclear. It has been proposed that positional information might be involved in the determination of cell fate (Grossniklaus and Schneitz, 1998; Sundaresan and Alandete-Saez, 2010; Sprunck and Gross-Hardt, 2011; Lituiev and Grossniklaus, 2014). During the syncytial phase, nuclei migrate and occupy predefined positions in the female gametophyte. In mutants with supernumerary nuclei, the position of the nuclei along the micropylar-chalazal axis of the embryo sac affects their cell fate, indicating that they are influenced by positional information (Table 3; Gross-Hard et al., 2007; Pagnussat et al., 2007; Moll et al., 2008a; Moll et al., 2008b; Johnston et al., 2010).

The role of auxin and cytokinin in establishing and maintaining polarity

It has been proposed that the plant hormone auxin plays a pivotal role in establishing and maintaining polarity by forming a gradient in the developing embryo sac (Pagnussat et al., 2009). The auxin gradient was thought to be mediated by auxin influx from sporophytic tissues at early stages and by localized biosynthesis at later stages of female gametophyte development. Abolishing the auxin gradient, by expressing the YUCCA1 (YUC1) auxin biosynthetic protein in the entire embryo sac or by modulating the auxin response by downregulating selected *AUXIN RESPONSE FACTOR* (*ARF* genes), led to the loss or, at low frequencies, the mis-expression of cell fate markers in the female gametophyte (Table 3; Pagnussat et al., 2009). However, theoretical models attempting to describe the auxin gradient in the female gametophyte showed that only very shallow auxin gradients can be established even when using the most favourable parameters (Lituiev et al., 2013). A sensitivity analysis demonstrated that the steepness of the obtained gradients is not sufficient to determine distinct cell fates (Lituiev et al., 2013). Furthermore, the reinvestigation of auxin signalling using various sensors failed to detect auxin in the female gametophyte but instead found auxin signalling to be restricted to the surrounding ovule tissues in a dynamic polar pattern (Ceccato et al., 2013; Lituiev et al., 2013). This polar auxin pattern in sporophytic tissues may non-cell-autonomously influence cell specification in the female gametophyte and may have been affected by manipulating the expression of *YUC1* and *ARF*s (Lituiev et al., 2013).

Auxin signalling is interrelated with the cytokinin pathway (Müller and Sheen, 2008; Bencivenga et al., 2012, Cheng et al., 2013) and, not surprisingly therefore, cytokinin signalling has also been shown to play a role in germline development. For example, cytokinin levels influence ovule patterning by affecting the

Development (2015) 142, 229-241 doi:10.1242/dev.102103

Table 3. Mutations involved in polarity and cell fate determination in the *Arabidopsis thaliana* female gametophyte

Mutation	Description	Phenotype	Reference
cki1	Mutation in a gene causing cytokinin-independent activation of the cytokinin signalling pathway	Arrest starting from FG4	Pischke et al., 2002; Hejátko et al., 2003
agp18	RNAi targeting *ARABINOGALACTAN PROTEIN18* transcripts	Arrest at FG1	Acosta-García and Vielle-Calzada, 2004
rbr1	Mutation in a core cell cycle regulator gene	Nuclear overproliferation	Ebel et al., 2004; Johnston et al., 2010
lis	Mutation in a gene encoding a component of the RNA splicing machinery	Synergids and the central cell adopt an egg cell-like fate	Gross-Hardt et al., 2007; Völz et al., 2012
eostre	Mutation leading to the misexpression of *BLH1*	One synergid cell differentiates into an additional egg cell	Pagnussat et al., 2007
clo/gfa1	Mutation in a gene encoding a component of the RNA splicing machinery	Synergids and the central cell adopt an egg cell-like fate	Moll et al., 2008b
ato	Mutation in a gene encoding a component of the RNA splicing machinery	Synergids and the central cell adopt an egg cell-like fate	Moll et al., 2008b
amiR-ARFa	amiRNA targeting transcripts of the auxin signalling pathway	Synergid identity lost, partly adopting an egg cell-like fate	Pagnussat et al., 2009
ahk2-7, ahk3-3, cre1-12	Triple mutant in genes encoding components of the cytokinin signalling pathway	Arrest at FG1	Cheng et al., 2013
ahp1, ahp2-1, ahp3, ahp4, ahp5	Multiple mutant in genes encoding components of the cytokinin signalling pathway	Arrest at FG7	Cheng et al., 2013
hda7	Mutation in a histone deacetylase gene	Arrest at FG4	Cigliano et al., 2013
myb64, myb119	Double mutant in MYB transcription factor genes	Arrest during FG5 transition	Rabiger and Drews, 2013

amiRNA, artificial microRNA; *BLH1*, *BELL1-LIKE HOMEODOMAIN1*; RNAi, RNA interference.
Female gametophyte stages (FG) are according to Christensen et al. (1997).

expression of *PIN-FORMED1* (*PIN1*), which encodes an auxin efflux carrier (Bencivenga et al., 2012; Luschnig and Vert, 2014). This is consistent with the finding that cytokinin regulates *PIN1* expression in roots (Dello Ioio et al., 2008; Ruzicka et al., 2009). In ovules, the regulatory pathway involves two transcription factors, SPOROCYTLESS/NOZZLE (SPL/NZZ) and the homeodomain protein BELL1 (BEL1). Mutations in the gene encoding SPL/NZZ, which is required for the initiation of megasporogenesis, lead to reduced expression of *PIN1*, while the effects of exogenous cytokinin are mediated by BEL1, which is important for ovule identity, and lead to an altered pattern of auxin signalling in the ovule (Table 3; Schiefthaler et al., 1999; Yang et al., 1999; Balasubramanian and Schneitz, 2000; Sieber et al., 2004; Brambilla et al., 2007; Bencivenga et al., 2012). In addition to *PIN1*, *PIN3* is expressed during ovule development (Ceccato et al., 2013). However, no effect on ovule or female germline development has been reported in *pin3* mutants and a potential functional interaction between cytokinin and *PIN3* has not been investigated (Ceccato et al., 2013). Thus, although an auxin gradient in the embryo sac could not be confirmed, auxin and cytokinin do play important roles in the sporophytic tissues of the ovule.

Consistent with the crosstalk between the auxin and cytokinin pathways, cytokinin is involved in communication between sporophytic ovule tissues and the developing female gametophyte (Table 3; Cheng et al., 2013). In *Arabidopsis*, different cytokinin receptors expressed in the chalazal ovule tissues act redundantly to regulate FMS specification (Table 3; Cheng et al., 2013). During megagametogenesis, the histidine protein kinase CYTOKININ-INDEPENDENT1 (CKI1) has important functions, and *cki1* mutants affect the mitotic divisions during gametophyte development (Table 3; Pischke et al., 2002; Hejátko et al., 2003; Cheng et al., 2013). Although related to the *Arabidopsis* histidine kinases (AHKs) AHK2, AHK3 and AHK4, which act as cytokinin receptors, CKI1 lacks a cytokinin-binding domain and activates the cytokinin signalling pathway in the absence of cytokinin (Kakimoto, 1996; Nakamura et al., 1999; Urao et al., 2000; Hwang and Sheen, 2001; Yamada et al., 2001; Mähönen et al., 2006). *Arabidopsis* double mutants affecting *MYB-DOMAIN PROTEIN64* (*MYB64*) and *MYB119*, which encode two closely related R2R3-MYB domain transcription factors, also display a *cki1*-like phenotype (Table 3; Rabiger and Drews, 2013). Double mutant *myb64 myb119* gametophytes undergo extra mitotic division cycles and usually fail to cellularize (Rabiger and Drews, 2013). In the few cellularized mutant embryo sacs, cell fate is not properly established and the polarity of the embryo sac is affected (Rabiger and Drews, 2013). Furthermore, while *MYB64* and *MYB119* act redundantly during female gametophyte development, *MYB119* but not *MYB64* is regulated by *CKI1* (Rabiger and Drews, 2013).

Epigenetic and post-transcriptional regulation of gametophyte development

In addition to hormonal pathways and transcription factors, epigenetic regulators are involved in establishing polarity in the developing gametophyte. Recently, a role for HISTONE DEACETYLASE7 (HDA7) during megagametogenesis and embryo development has been demonstrated (Cigliano et al., 2013). In *hda7-2* mutants at the four-nucleate stage of megagametogenesis, the two nuclei located at the micropylar pole degenerate, suggesting that histone deacetylation is required for survival and possibly for fate determination of the micropylar nuclei (Table 3; Cigliano et al., 2013). Other important gene regulatory control mechanisms involve the storage of mRNAs in mRNA-protein complexes, mRNA processing and mRNA degradation (reviewed by Hafidh et al., 2011). Regulation of the asymmetric distribution and processing of mRNAs involving RNA-binding proteins is known to be a determinant of protein gradients, cell polarity, cell fate decisions and patterning during development (Hafidh et al., 2011). For example, this is well described in *Drosophila* embryo genesis but also relevant for polar pollen tube growth in plants (Hafidh et al., 2011). In agreement with the emerging roles of mRNA storage and processing, components of the RNA splicing machinery have been identified as being crucial for cell type specification and the restriction

of gametic fate in the *Arabidopsis* embryo sac (Table 3; Gross-Hardt et al., 2007: Moll et al., 2008b; Völz et al., 2012). In *lachesis* (*lis*) mutants, the expression of a marker for egg cell identity extends to adjacent gametophytic cells, the synergids and the central cell (Table 3; Gross-Hardt et al., 2007). As the phenotype becomes stronger as time progresses, LIS may predominantly play a role in maintaining egg cell identity. *LIS* encodes a homologue of the yeast splicing factor PRP4 (Gross-Hardt et al., 2007). Similar to *LIS*, *CLOTHO*/*GAMETOPHYTIC FACTOR1* (*CLO*/*GFA1*) and *ATROPUS* (*ATO*) are also important for restricting gametic fate in the mature gametophyte (Table 3; Moll et al., 2008b). *CLO*/*GFA1* encodes a homologue of Snu114, an essential component of the spliceosome, while *ATO* encodes the *Arabidopsis* homologue of SF3a60, which plays a role in pre-spliceosome formation (Moll et al., 2008b). The activities of *LIS* and *CLO* are related, as *CLO* is important for the tissue specificity of *LIS* expression (Moll et al., 2008b). *LIS* is strongly enriched in female gametes, suggesting that it regulates the maintenance of cell fate by lateral inhibition of the adjacent accessory cells in the female gametophyte, the synergid and antipodal cells (Gross-Hardt et al., 2007; Moll et al., 2008b; Völz et al., 2012). In summary, the splicing machinery is important for the specification and maintenance of cellular identity in the female gametophyte. Whether this is mediated through specific effects of some of its components in the embryo sac or caused by a general deficiency in splicing – also affecting pre-mRNAs of cell specification factors – remains to be determined.

In conclusion, although many mutants that exhibit disrupted embryo sac polarity or cell type-specific expression have been identified over the past decade, we are still far from understanding these processes at the molecular level. Currently, we have a partial list of components involved in cell specification but we do not understand how they work together to pattern the female gametophyte. Importantly, many of the observed effects may be indirect, e.g. caused by the mis-positioning of nuclei in the embryo sac, or the identified factors act after the initial specification of cell fate, in the maintenance of cell identity or during cell differentiation. A clear candidate for a cell fate determinant – one that cell-autonomously specifies cell type identity – is still being sought after. Transcription factors of the RKD family can at least partially reprogramme sporophytic cells towards an egg cell fate when overexpressed (Koszegi et al., 2011) but, owing to genetic redundancy, functional analyses of these transcription factors proved difficult and their potential gametophytic phenotypes are unknown.

Conclusions

The male and female plant germlines are ideal models for studying the role of polarity, cell specification processes and the transition from sporophytic to gametophytic fate, which is a key step in the plant life cycle. Apart from being scientifically fascinating, understanding the molecular mechanisms underlying the specification and development of plant reproductive lineages is relevant for targeted manipulations of reproduction for crop improvement and seed production, in particular to achieve the longstanding goal of engineering apomixis in crop plants. Important aspects will be to determine whether distinct or similar genetic and epigenetic modifications govern apomixis in different species, involving aposporous and diplosporous accessions, and to identify common features.

Recent studies have yielded important insights into various aspects of sexual and apomictic germline formation, providing a glimpse of the complex regulatory mechanisms required to control reproductive development in plants. In this Review, we have discussed studies that address the genetic basis underlying the transition from somatic to germline fate, the repression of additional reproductive lineages, and cell specification during megagametogenesis, which together ensure reproductive success. As highlighted above, many different pathways are involved, including hormonal pathways, epigenetic regulation via small RNAs, transcriptional regulation by transcription factors and post-transcriptional control mechanisms. However, we currently do not know how these pathways are interconnected to form regulatory networks. Building on recent investigations, an important aim of future research will be to compare whole transcriptome analyses and combine these with detailed studies of the molecular mechanisms at play in different species, while taking evolutionary aspects into consideration.

Finally, little is known about the influence of stress or changing environmental conditions on germline specification. Interestingly, heat stress as well as the abundance of ROS were reported to influence the meiotic versus apomeiotic fate decision of spore mother cells. However, contradicting hypotheses have been portrayed, highlighting a need for more investigations in this area. In the longer term, this may be important not only for a better understanding of the regulatory networks underlying reproductive development and their interactions with environmental factors, but also for the improvement of agricultural plants under changing climate conditions.

Acknowledgements

We thank our colleagues in the Grossniklaus laboratory for interesting discussions and three anonymous reviewers for their helpful comments, which helped us to improve the manuscript.

Competing interests

The authors declare no competing financial interests.

Funding

Work on gametophyte development, apomixis and epigenetic gene regulation in U.G.'s laboratory is supported by the University of Zürich, and by grants from the 'Staatssekretariat für Bildung und Forschung' in the framework of COST action FA0903 (to U.G. and A.S.), the Swiss National Science Foundation (to U.G.) and the European Research Council (to U.G.).

References

Abiko, M., Maeda, H., Tamura, K., Hara-Nishimura, I. and Okamoto, T. (2013). Gene expression profiles in rice gametes and zygotes: identification of gamete-enriched genes and up- or down-regulated genes in zygotes after fertilization. *J. Exp. Bot.* **64**, 1927-1940.

Acosta-García, G. and Vielle-Calzada, J.-P. (2004). A classical arabinogalactan protein is essential for the initiation of female gametogenesis in *Arabidopsis*. *Plant Cell* **16**, 2614-2628.

Albrecht, C., Russinova, E., Hecht, V., Baaijens, E. and de Vries, S. (2005). The *Arabidopsis thaliana* SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASES 1 and 2 control male sporogenesis. *Plant Cell* **17**, 3337-3349.

Anderson, S. N., Johnson, C. S., Jones, D. S., Conrad, L. J., Gou, X., Russell, S. D. and Sundaresan, V. (2013). Transcriptomes of isolated *Oryza sativa* gametes characterized by deep sequencing: evidence for distinct sex-dependent chromatin and epigenetic states before fertilization. *Plant J.* **76**, 729-741.

Asker, S. and Jerling, E. L. (1992). *Apomixis in Plants*. Boca Raton, Florida, USA: CRC Press.

Bachelier, J. B. and Friedman, W. E. (2011). Female gamete competition in an ancient angiosperm lineage. *Proc. Natl. Acad. Sci. USA* **108**, 12360-12365.

Balasubramanian, S. and Schneitz, K. (2000). *NOZZLE* regulates proximal-distal pattern formation, cell proliferation and early sporogenesis during ovule development in *Arabidopsis thaliana*. *Development* **127**, 4227-4238.

Bencivenga, S., Simonini, S., Benková, E. and Colombo, L. (2012). The transcription factors BEL1 and SPL are required for cytokinin and auxin signaling during ovule development in *Arabidopsis*. *Plant Cell* **24**, 2886-2897.

Berger, F. and Twell, D. (2011). Germline specification and function in plants. *Annu. Rev. Plant Biol.* **62**, 461-484.

Bicknell, R. A. and Koltunow, A. M. (2004). Understanding apomixis: recent advances and remaining conundrums. *Plant Cell* **16** Suppl. 1, S228-S245.

Boavida, L. C., Becker, J. D. and Feijo, J. A. (2005). The making of gametes in higher plants. *Int. J. Dev. Biol.* **49**, 595-614.

Borg, M., Brownfield, L. and Twell, D. (2009). Male gametophyte development: a molecular perspective. *J. Exp. Bot.* **60**, 1465-1478.

Borges, F., Pereira, P. A., Slotkin, R. K., Martienssen, R. A. and Becker, J. D. (2011). MicroRNA activity in the *Arabidopsis* male germline. *J. Exp. Bot.* **62**, 1611-1620.
Borner, G. H. H., Lilley, K. S., Stevens, T. J. and Dupree, P. (2003). Identification of glycosylphosphatidylinositol-anchored proteins in *Arabidopsis*. A proteomic and genomic analysis. *Plant Physiol.* **132**, 568-577.
Brambilla, V., Battaglia, R., Colombo, M., Masiero, S., Bencivenga, S., Kater, M. M. and Colombo, L. (2007). Genetic and molecular interactions between BELL1 and MADS box factors support ovule development in *Arabidopsis*. *Plant Cell* **19**, 2544-2556.
Brownfield, L. and Köhler, C. (2011). Unreduced gamete formation in plants: mechanisms and prospects. *J. Exp. Bot.* **62**, 1659-1668.
Canales, C., Bhatt, A. M., Scott, R. and Dickinson, H. (2002). *EXS*, a putative LRR receptor kinase, regulates male germline cell number and tapetal identity and promotes seed development in *Arabidopsis*. *Curr. Biol.* **12**, 1718-1727.
Ceccato, L., Masiero, S., Sinha Roy, D., Bencivenga, S., Roig-Villanova, I., Ditengou, F. A., Palme, K., Simon, R. and Colombo, L. (2013). Maternal control of *PIN1* is required for female gametophyte development in *Arabidopsis*. *PLoS ONE* **8**, e66148.
Cheng, C.-Y., Mathews, D. E., Schaller, G. E. and Kieber, J. J. (2013). Cytokinin-dependent specification of the functional megaspore in the *Arabidopsis* female gametophyte. *Plant J.* **73**, 929-940.
Chettoor, A. M., Givan, S. A., Cole, R. A., Coker, C. T., Unger-Wallace, E., Vejlupkova, Z., Vollbrecht, E., Fowler, J. E. and Evans, M. M. S. (2014). Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. *Genome Biol.* **15**, 414.
Christensen, C. A., King, E. J., Jordan, J. R. and Drews, G. N. (1997). Megagametogenesis in *Arabidopsis* wild type and the *Gf* mutant. *Sex. Plant Reprod.* **10**, 49-64.
Cigliano, R. A., Cremona, G., Paparo, R., Termolino, P., Perrella, G., Gutzat, R., Consiglio, M. F. and Conicella, C. (2013). Histone deacetylase AtHDA7 is required for female gametophyte and embryo development in *Arabidopsis*. *Plant Physiol.* **163**, 431-440.
Clark, J. P. and Lau, N. C. (2014). Piwi proteins and piRNAs step onto the systems biology stage. *Adv. Exp. Med. Biol.* **825**, 159-197.
Colcombet, J., Boisson-Dernier, A., Ros-Palau, R., Vera, C. E. and Schroeder, J. I. (2005). *Arabidopsis* SOMATIC EMBRYOGENESIS RECEPTOR KINASES1 and 2 are essential for tapetum development and microspore maturation. *Plant Cell* **17**, 3350-3361.
Crismani, W., Girard, C. and Mercier, R. (2013). Tinkering with meiosis. *J. Exp. Bot.* **64**, 55-65.
De Storme, N. and Geelen, D. (2013). Sexual polyploidization in plants - cytological mechanisms and molecular regulation. *New Phytol.* **198**, 670-684.
De Storme, N. and Geelen, D. (2014). The impact of environmental stress on male reproductive development in plants: biological processes and molecular mechanisms. *Plant Cell Environ.* **37**, 1-18.
Dello Ioio, R., Nakamura, K., Moubayidin, L., Perilli, S., Taniguchi, M., Morita, M. T., Aoyama, T., Costantino, P. and Sabatini, S. (2008). A genetic framework for the control of cell division and differentiation in the root meristem. *Science* **322**, 1380-1384.
Demesa-Arévalo, E. and Vielle-Calzada, J.-P. (2013). The classical arabinogalactan protein AGP18 mediates megaspore selection in *Arabidopsis*. *Plant Cell* **25**, 1274-1287.
Drews, G. N. and Koltunow, A. M. G. (2011). The female gametophyte. *Arabidopsis Book* **9**, e0155.
d'Erfurth, I., Jolivet, S., Froger, N., Catrice, O., Novatchkova, M. and Mercier, R. (2009). Turning meiosis into mitosis. *PLoS Biol.* **7**, e1000124.
d'Erfurth, I., Cromer, L., Jolivet, S., Girard, C., Horlow, C., Sun, Y., To, J. P. C., Berchowitz, L. E., Copenhaver, G. P. and Mercier, R. (2010). The cyclin-A CYCA1;2/TAM is required for the meiosis I to meiosis II transition and cooperates with OSD1 for the prophase to first meiotic division transition. *PLoS Genet.* **6**, e1000989.
Ebel, C., Mariconti, L. and Gruissem, W. (2004). Plant retinoblastoma homologues control nuclear proliferation in the female gametophyte. *Nature* **429**, 776-780.
Ellis, M., Egelund, J., Schultz, C. J. and Bacic, A. (2010). Arabinogalactan-proteins: key regulators at the cell surface? *Plant Physiol.* **153**, 403-419.
Feng, X. and Dickinson, H. G. (2010). Tapetal cell fate, lineage and proliferation in the *Arabidopsis* anther. *Development* **137**, 2409-2416.
Garcia-Aguilar, M., Michaud, C., Leblanc, O. and Grimanelli, D. (2010). Inactivation of a DNA methylation pathway in maize reproductive organs results in apomixis-like phenotypes. *Plant Cell* **22**, 3249-3267.
Germanà, M. A. (2011). Gametic embryogenesis and haploid technology as valuable support to plant breeding. *Plant Cell Rep.* **30**, 839-857.
Grimanelli, D., Leblanc, O., Perotti, E. and Grossniklaus, U. (2001). Developmental genetics of gametophytic apomixis. *Trends Genet.* **17**, 597-604.
Gross-Hardt, R., Kägi, C., Baumann, N., Moore, J. M., Baskar, R., Gagliano, W. B., Jürgens, G. and Grossniklaus, U. (2007). *LACHESIS* restricts gametic cell fate in the female gametophyte of *Arabidopsis*. *PLoS Biol.* **5**, e47.
Grossniklaus, U. (2001). From sexuality to apomixis: molecular and genetic approaches. In *The Flowering of Apomixis: From Mechanisms to Genetic Engineering* (ed. Y. Savidan, J. Carman and T. Dresselhaus), pp. 168-211. Mexico, DF: CIMMYT.
Grossniklaus, U. (2011). Plant germline development: a tale of cross-talk, signaling, and cellular interactions. *Sex. Plant Reprod.* **24**, 91-95.
Grossniklaus, U. and Schneitz, K. (1998). The molecular and genetic basis of ovule and megagametophyte development. *Semin. Cell Dev. Biol.* **9**, 227-238.
Grossniklaus, U., Koltunow, A. and van Lookeren Campagne, M. (1998a). A bright future for apomixis. *Trends Plant Sci.* **3**, 415-416.
Grossniklaus, U., Moore, J. M. and Gagliano, W. B. (1998b). Molecular and genetic approaches to understanding and engineering apomixis: *Arabidopsis* as a powerful tool. In *Advances in Hybrid Rice Technology. Proceedings of the 3rd International Symposium on Hybrid Rice 1996* (ed. S. S. Virmani, E. A. Siddiq and K. Muralidharan), pp. 187-211. Manila, Philippines: International Rice Research Institute.
Gutierrez-Marcos, J. F. and Dickinson, H. G. (2012). Epigenetic reprogramming in plant reproductive lineages. *Plant Cell Physiol.* **53**, 817-823.
Hafidh, S., Čapková, V. and Honys, D. (2011). Safe keeping the message: mRNP complexes tweaking after transcription. *Adv. Exp. Med. Biol.* **722**, 118-136.
Haig, D. (1990). New perspectives on the angiosperm female gametophyte. *Botanical Rev.* **56**, 236-274.
Hejátko, J., Pernisová, M., Eneva, T., Palme, K. and Brzobohatý, B. (2003). The putative sensor histidine kinase CKI1 is involved in female gametophyte development in *Arabidopsis*. *Mol. Genet. Genomics* **269**, 443-453.
Hörandl, E. and Hadacek, F. (2013). The oxidative damage initiation hypothesis for meiosis. *Plant Reprod.* **26**, 351-367.
Huang, B.-Q. and Russell, S. D. (1992). Female germ unit: organization, isolation, and function. *Int. Rev. Cytol.* **140**, 233-293.
Huang, B. Q. and Sheridan, W. F. (1996). Embryo sac development in the maize *indeterminate gametophyte1* mutant: abnormal nuclear behavior and defective microtubule organization. *Plant Cell* **8**, 1391-1407.
Hwang, I. and Sheen, J. (2001). Two-component circuitry in *Arabidopsis* cytokinin signal transduction. *Nature* **413**, 383-389.
Jia, G., Liu, X., Owen, H. A. and Zhao, D. (2008). Signaling of cell fate determination by the TPD1 small protein and EMS1 receptor kinase. *Proc. Natl. Acad. Sci. USA* **105**, 2220-2225.
Johnston, A. J., Kirioukhova, O., Barrell, P. J., Rutten, T., Moore, J. M., Baskar, R., Grossniklaus, U. and Gruissem, W. (2010). Dosage-sensitive function of *RETINOBLASTOMA RELATED* and convergent epigenetic control are required during the *Arabidopsis* life cycle. *PLoS Genet.* **6**, pe1000988.
Kakimoto, T. (1996). CKI1, a histidine kinase homolog implicated in cytokinin signal transduction. *Science* **274**, 982-985.
Kelliher, T. and Walbot, V. (2012). Hypoxia triggers meiotic fate acquisition in maize. *Science* **337**, 345-348.
Kelliher, T. and Walbot, V. (2014). Maize germinal cell initials accommodate hypoxia and precociously express meiotic genes. *Plant J.* **77**, 639-652.
Koltunow, A. M. (1993). Apomixis: embryo sacs and embryos formed without meiosis or fertilization in ovules. *Plant Cell* **5**, 1425-1437.
Koltunow, A. and Grossniklaus, U. (2003). Apomixis: a developmental perspective. *Annu. Rev. Plant Biol.* **54**, 547-574.
Koltunow, A. M., Bicknell, R. A. and Chaudhury, A. M. (1995). Apomixis: molecular strategies for the generation of genetically identical seeds without fertilization. *Plant Physiol.* **108**, 1345-1352.
Koltunow, A. M. G., Johnson, S. D., Rodrigues, J. C. M., Okada, T., Hu, Y., Tsuchiya, T., Wilson, S., Fletcher, P., Ito, K., Suzuki, G. et al. (2011). Sexual reproduction is the default mode in apomictic *Hieracium* subgenus *Pilosella*, in which two dominant loci function to enable apomixis. *Plant J.* **66**, 890-902.
Koszegi, D., Johnston, A. J., Rutten, T., Czihal, A., Altschmied, L., Kumlehn, J., Wüst, S. E. J., Kirioukhova, O., Gheyselinck, J., Grossniklaus, U. et al. (2011). Members of the RKD transcription factor family induce an egg cell-like gene expression program. *Plant J.* **67**, 280-291.
Leblanc, O., Armstead, I., Pessino, S., Ortiz, J. P. A., Evans, C., doValle, C. and Hayward, M. D. (1997). Non-radioactive mRNA fingerprinting to visualise gene expression in mature ovaries of *Brachiaria* hybrids derived from *B.* brizantha, an apomictic tropical forage. *Plant Sci.* **126**, 49-58.
Lin, B.-Y. (1984). Ploidy bareer to endosperm development in maize. *Genetics* **107**, 103-115.
Lituiev, D. S. and Grossniklaus, U. (2014). Patterning of the angiosperm female gametophyte through the prism of theoretical paradigms. *Biochem. Soc. Trans.* **42**, 332-339.
Lituiev, D. S., Krohn, N. G., Müller, B., Jackson, D., Hellriegel, B., Dresselhaus, T. and Grossniklaus, U. (2013). Theoretical and experimental evidence indicates that there is no detectable auxin gradient in the angiosperm female gametophyte. *Development* **140**, 4544-4553.
Luschnig, C. and Vert, G. (2014). The dynamics of plant plasma membrane proteins: PINs and beyond. *Development* **141**, 2924-2938.
Maheshwari, P. (1950). *An Introduction to the Embryology of Angiosperms*. New York: McGraw-Hill Publications.

Mähönen, A. P., Higuchi, M., Törmäkangas, K., Miyawaki, K., Pischke, M. S., Sussman, M. R., Helariutta, Y. and Kakimoto, T. (2006). Cytokinins regulate a bidirectional phosphorelay network in *Arabidopsis*. *Curr. Biol.* **16**, 1116-1122.
Mallory, A. and Vaucheret, H. (2010). Form, function, and regulation of ARGONAUTE proteins. *Plant Cell.* **22**, 3879-3889.
Marimuthu, M. P. A., Jolivet, S., Ravi, M., Pereira, L., Davda, J. N., Cromer, L., Wang, L., Nogué, F., Chan, S. W. L., Siddiqi, I. et al. (2011). Synthetic clonal reproduction through seeds. *Science* **331**, 876.
Meister, G. (2013). Argonaute proteins: functional insights and emerging roles. *Nat. Rev. Genet.* **14**, 447-459.
Moll, C., Nielsen, N. and Gross-Hardt, R. (2008a). Mutants with aberrant numbers of gametic cells shed new light on old questions. *Plant Biol. (Stuttg)* **10**, 529-533.
Moll, C. von Lyncker, L., Zimmermann, S., Kägi, C., Baumann, N., Twell, D., Grossniklaus, U. and Gross-Hardt, R. (2008b). *CLO*/*GFA1* and *ATO* are novel regulators of gametic cell fate in plants. *Plant J.* **56**, 913-921.
Müller, B., and Sheen, J. (2008). Cytokinin and auxin iteraction in root stem-cell specification during early embryogenesis. *Nature* **453**, 1094-1097.
Nakamura, A., Kakimoto, T., Imamura, A., Suzuki, T., Ueguchi, C. and Mizuno, T. (1999). Biochemical characterization of a putative cytokinin-responsive His-kinase, CKI1, from *Arabidopsis thaliana*. *Biosci. Biotechnol. Biochem.* **63**, 1627-1630.
Nonomura, K.-I., Miyoshi, K., Eiguchi, M., Suzuki, T., Miyao, A., Hirochika, H. and Kurata, N. (2003). The *MSP1* gene is necessary to restrict the number of cells entering into male and female sporogenesis and to initiate anther wall formation in rice. *Plant Cell* **15**, 1728-1739.
Nonomura, K.-I., Morohoshi, A., Nakano, M., Eiguchi, M., Miyao, A., Hirochika, H. and Kurata, N. (2007). A germ cell specific gene of the *ARGONAUTE* family is essential for the progression of premeiotic mitosis and meiosis during sporogenesis in rice. *Plant Cell* **19**, 2583-2594.
Okada, T., Hu, Y., Tucker, M. R., Taylor, J. M., Johnson, S. D., Spriggs, A., Tsuchiya, T., Oelkers, K., Rodrigues, J. C. M. and Koltunow, A. M. (2013). Enlarging cells initiating apomixis in *Hieracium praealtum* transition to an embryo sac program prior to entering mitosis. *Plant Physiol.* **163**, 216-231.
Olmedo-Monfil, V., Durán-Figueroa, N., Arteaga-Vázquez, M., Demesa-Arévalo, E., Autran, D., Grimanelli, D., Slotkin, R. K., Martienssen, R. A. and Vielle-Calzada, J.-P. (2010). Control of female gamete formation by a small RNA pathway in *Arabidopsis*. *Nature* **464**, 628-632.
Pagnussat, G. C., Yu, H.-J. and Sundaresan, V. (2007). Cell-fate switch of synergid to egg cell in *Arabidopsis eostre* mutant embryo sacs arises from misexpression of the BEL1-like homeodomain gene *BLH1*. *Plant Cell* **19**, 3578-3592.
Pagnussat, G. C., Alandete-Saez, M., Bowman, J. L. and Sundaresan, V. (2009). Auxin-dependent patterning and gamete specification in the *Arabidopsis* female gametophyte. *Science* **324**, 1684-1689.
Pecrix, Y., Rallo, G., Folzer, H., Cigna, M., Gudin, S. and Le Bris, M. (2011). Polyploidization mechanisms: temperature environment can induce diploid gamete formation in *Rosa sp*. *J. Exp. Bot.* **62**, 3587-3597.
Pischke, M. S., Jones, L. G., Otsuga, D., Fernandez, D. E., Drews, G. N. and Sussman, M. R. (2002). An *Arabidopsis* histidine kinase is essential for megagametogenesis. *Proc. Natl. Acad. Sci. USA* **99**, 15800-15805.
Rabiger, D. S. and Drews, G. N. (2013). MYB64 and MYB119 are required for cellularization and differentiation during female gametogenesis in *Arabidopsis thaliana*. *PLoS Genet.* **9**, pe1003783.
Raghavan, V. (1997). *Molecular Embryology of Flowering Plants*. Cambridge, UK: Cambridge University Press.
Ravi, M., Marimuthu, M. P. A. and Siddiqi, I. (2008). Gamete formation without meiosis in *Arabidopsis*. *Nature* **451**, 1121-1124.
Russell, S. D. (1979). Fine structure of megagametophyte development in *Zea mays*. *Can. J. Bot.* **57**, 1093-1110.
Ruzicka, K., Simásková, M., Duclercq, J., Petrásek, J., Zazímalová, E., Simon, S., Friml, J., Van Montagu, M. C. E. and Benková, E. (2009). Cytokinin regulates root meristem activity via modulation of the polar auxin transport. *Proc. Natl. Acad. Sci. USA* **106**, 4284-4289.
Schiefthaler, U., Balasubramanian, S., Sieber, P., Chevalier, D., Wisman, E. and Schneitz, K. (1999). Molecular analysis of *NOZZLE*, a gene involved in pattern formation and early sporogenesis during sex organ development in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **96**, 11664-11669.
Schmid, M. W., Schmidt, A., Klostermeier, U. C., Barann, M., Rosenstiel, P. and Grossniklaus, U. (2012). A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing. *PLoS ONE* **7**, pe29685.
Schmidt, A., Wüest, S. E., Vijverberg, K., Baroux, C., Kleen, D. and Grossniklaus, U. (2011). Transcriptome analysis of the *Arabidopsis* megaspore mother cell uncovers the importance of RNA helicases for plant germline development. *PLoS Biol.* **9**, e1001155.
Schmidt, A., Schmid, M. W. and Grossniklaus, U. (2012). Analysis of plant germline development by high-throughput RNA profiling: technical advances and new insights. *Plant J.* **70**, 18-29.
Schmidt, A., Schmid, M. W., Klostermeier, U. C., Qi, W., Guthörl, D., Sailer, C., Waller, M., Rosenstiel, P. and Grossniklaus, U. (2014). Apomictic and sexual germline development differ with respect to cell cycle, transcriptional, hormonal and epigenetic regulation. *PLoS Genet.* **10**, e1004476.
Seifert, G. J. and Roberts, K. (2007). The biology of arabinogalactan proteins. *Annu. Rev. Plant Biol.* **58**, 137-161.
Sharbel, T. F., Voigt, M. L., Corral, J. M., Thiel, T., Varshney, A., Kumlehn, J., Vogel, H. and Rotter, B. (2009). Molecular signatures of apomictic and sexual ovules in the *Boechera holboellii* complex. *Plant J.* **104**, 14026-14031.
Sharbel, T. F., Voigt, M. L., Corral, J. M., Galla, G., Kumlehn, J., Klukas, C., Schreiber, F., Vogel, H. and Rotter, B. (2010). Apomictic and sexual ovules of *Boechera* display heterochronic global gene expression patterns. *Plant Cell* **22**, 655-671.
She, W., Grimanelli, D., Rutowicz, K., Whitehead, M. W. J., Puzio, M., Kotlinski, M., Jerzmanowski, A. and Baroux, C. (2013). Chromatin reprogramming during the somatic-to-reproductive cell fate transition in plants. *Development* **140**, 4008-4019.
Sheridan, W. F., Avalkina, N. A., Shamrov, I. I., Batygina, T. B. and Golubovskaya, I. N. (1996). The *mac1* gene: controlling the commitment to the meiotic pathway in maize. *Genetics* **142**, 1009-1020.
Sheridan, W. F., Golubeva, E. A., Abrhamova, L. I. and Golubovskaya, I. N. (1999). The *mac1* mutation alters the developmental fate of the hypodermal cells and their cellular progeny in the maize anther. *Genetics* **153**, 933-941.
Sieber, P., Petrascheck, M., Barberis, A. and Schneitz, K. (2004). Organ polarity in *Arabidopsis*. NOZZLE physically interacts with members of the YABBY family. *Plant Physiol.* **135**, 2172-2185.
Singh, M., Goel, S., Meeley, R. B., Dantec, C., Parrinello, H., Michaud, C., Leblanc, O. and Grimanelli, D. (2011). Production of viable gametes without meiosis in maize deficient for an ARGONAUTE protein. *Plant Cell* **23**, 443-458.
Spillane, C., Steimer, A. and Grossniklaus, U. (2001). Apomixis in agriculture: the quest for clonal seeds. *Sex. Plant Reprod.* **14**, 179-187.
Spillane, C., Curtis, M. D. and Grossniklaus, U. (2004). Apomixis technology development-virgin births in farmers' fields? *Nat. Biotechnol.* **22**, 687-691.
Sprunck, S. and Gross-Hardt, R. (2011). Nuclear behavior, cell polarity, and cell specification in the female gametophyte. *Sex. Plant Reprod.* **24**, 123-136.
Sundaresan, V. and Alandete-Saez, M. (2010). Pattern formation in miniature: the female gametophyte of flowering plants. *Development* **137**, 179-189.
Tucker, M. R., Araujo, A.-C. G., Paech, N. A., Hecht, V., Schmidt, E. D. L., Rossell, J.-B., de Vries, S. C. and Koltunow, A. M. G. (2003). Sexual and apomictic reproduction in *Hieracium* subgenus *pilosella* are closely interrelated developmental pathways. *Plant Cell* **15**, 1524-1537.
Tucker, M. R., Okada, T., Hu, Y., Scholefield, A., Taylor, J. M. and Koltunow, A. M. G. (2012). Somatic small RNA pathways promote the mitotic events of megagametogenesis during female reproductive development in *Arabidopsis*. *Development* **139**, 1399-1404.
Twell, D. (2011). Male gametogenesis and germline specification in flowering plants. *Sex. Plant Reprod.* **24**, 149-160.
Urao, T., Miyata, S., Yamaguchi-Shinozaki, K. and Shinozaki, K. (2000). Possible His to Asp phosphorelay signaling in an *Arabidopsis* two-component system. *FEBS Lett.* **478**, 227-232.
Vielle-Calzada, J.-P., Crane, C. F. and Stelly, D. M. (1996). Apomixis–the asexual revolution. *Science* **274**, 1322-1323.
Völz, R., von Lyncker, L., Baumann, N., Dresselhaus, T., Sprunck, S. and Gross-Hardt, R. (2012). LACHESIS-dependent egg-cell signaling regulates the development of female gametophytic cells. *Development* **139**, 498-502.
Wang, C.-J. R., Nan, G.-L., Kelliher, T., Timofejeva, L., Vernoud, V., Golubovskaya, I. N., Harper, L., Egger, R., Walbot, V. and Cande, W. Z. (2012). Maize *multiple archesporial cells1* (*mac1*), an ortholog of rice *TDL1A*, modulates cell proliferation and identity in early anther development. *Development* **139**, 2594-2603.
Willemse, M. T. M. and van Went, J. L. (1984). The female gametophyte. In *Embryology of Angiosperms* (ed. B. M. Johri), pp. 159-196. Berlin: Springer-Verlag.
Wüest, S. E., Vijverberg, K., Schmidt, A., Weiss, M., Gheyselinck, J., Lohr, M., Wellmer, F., Rahnenführer, J., von Mering, C. and Grossniklaus, U. (2010). *Arabidopsis* female gametophyte gene expression map reveals similarities between plant and animal gametes. *Curr. Biol.* **20**, 506-512.
Wüest, S. E., Schmid, M. W. and Grossniklaus, U. (2013). Cell-specific expression profiling of rare cell types as exemplified by its impact on our understanding of female gametophyte development. *Curr. Opin. Plant Biol.* **16**, 41-49.
Yajima, M. and Wessel, G. M. (2011). The multiple hats of Vasa: its functions in the germline and in cell cycle progression. *Mol. Reprod. Dev.* **78**, 861-867.
Yamada, H., Suzuki, T., Terada, K., Takei, K., Ishikawa, K., Miwa, K., Yamashino, T. and Mizuno, T. (2001). The *Arabidopsis* AHK4 histidine kinase is a cytokinin-binding receptor that transduces cytokinin signals across the membrane. *Plant Cell. Physiol.* **42**, 1017-1023.
Yang, W.-C. and Sundaresan, V. (2000). Genetics of gametophyte biogenesis in *Arabidopsis*. *Curr. Opin. Plant Biol.* **3**, 53-57.

Yang, W.-C., Ye, D., Xu, J. and Sundaresan, V. (1999). The *SPOROCYTELESS* gene of *Arabidopsis* is required for initiation of sporogenesis and encodes a novel nuclear protein. *Genes Dev.* **13**, 2108-2117.

Yang, S.-L., Xie, L.-F., Mao, H.-Z., Puah, C. S., Yang, W.-C., Jiang, L., Sundaresan, V. and Ye, D. (2003). *TAPETUM DETERMINANT1* is required for cell specialization in the *Arabidopsis* anther. *Plant Cell* **15**, 2792-2804.

Yang, S.-L., Jiang, L., Puah, C. S., Xie, L.-F., Zhang, X.-Q., Chen, L.-Q., Yang, W.-C. and Ye, D. (2005). Overexpression of *TAPETUM DETERMINANT1* alters the cell fates in the *Arabidopsis* carpel and tapetum via genetic interaction with *excess microsporocytes1*/*extra sporogenous cells*. *Plant. Physiol.* **139**, 186-191.

Youl, J. J., Bacic, A. and Oxley, D. (1998). Arabinogalactan-proteins from *Nicotiana alata* and *Pyrus communis* contain glycosylphosphatidylinositol membrane anchors. *Proc. Natl. Acad. Sci. USA* **95**, 7921-7926.

Zhang, D. and Yang, L. (2014). Specification of tapetum and microsporocyte cells within the anther. *Curr. Opin. Plant Biol.* **17**, 49-55.

Zhang, Y., Yang, J. and Showalter, A. M. (2011). AtAGP18, a lysine-rich arabinogalactan protein in *Arabidopsis thaliana*, functions in plant growth and development as a putative co-receptor for signal transduction. *Plant Signal. Behav.* **6**, 855-857.

Zhao, D.-Z., Wang, G.-F., Speal, B. and Ma, H. (2002). The *excess microsporocytes1* gene encodes a putative leucine-rich repeat receptor protein kinase that controls somatic and reproductive cell fates in the *Arabidopsis* anther. *Genes Dev.* **16**, 2021-2031.

Zhao, X., de Palma, J., Oane, R., Gamuyao, R., Luo, M., Chaudhury, A., Hervé, P., Xue, Q. and Bennett, J. (2008). OsTDL1A binds to the LRR domain of rice receptor kinase MSP1, and is required to limit sporocyte numbers. *Plant J.* **54**, 375-387.